# Notes on the Large Data Set OCR Mathematics B (MEI) (H630/H640)

**LDS 8 used for:**

**H630/02 in June 2025**

**H640/02 in June 2026**

These notes outline the requirements for OCR’s large data set for specification B(MEI), include some notes on interesting features of the large data set and some links to some useful resources.

Please note that the approach outlined here and the suggested activities are suggestions only; you are free to deliver this in any way that suits you, your students and your setting.

If you have any comments on the contents of this guide, including suggestions for other activities, please get in touch with the Mathematics Subject Advisors at Maths@ocr.org.uk .

In the ‘Overview’ section we take a look at the *requirements* in terms of teaching from the Department for Education and in terms of assessment from Ofqual, and introduce one of the large data sets chosen for OCR’s Maths B (MEI) specification. We also suggest some software you will find useful when working with the LDS.

In the ‘Key Features’ section we will take a look at some important aspects of the structure and set up of the data. We will also explore some aspects of data cleaning in this section, including a list of things to look out for in the way that the data were presented by the London Datastore, along with some ways to explore these issues and to set up your students to stumble across them.

In ‘Online resources’ we highlight some useful places to go for more contextual information, maps, more data, tools and so on.

You can find teaching activities using the LDS within MEI’s Integral resources: <https://integralmaths.org/statistics.php>

***DISCLAIMER***

This resource was designed using the most up to date information from the specification at the time it was published. Specifications are updated over time, which means there may be contradictions between the resource and the specification, therefore please use the information on the latest specification at all times.If you do notice a discrepancy please contact us on the following email address: resources.feedback@ocr.org.uk

### Contents

[1 Overview 1](#_Toc137461615)

[1.1 Introduction 1](#_Toc137461616)

[1.2 MEI Large Data Set 8 (for AS in June 2025 and A Level in June 2026) 2](#_Toc137461617)

[1.3 Why three large data sets for the MEI specification? 3](#_Toc137461618)

[1.4 Features of the three large data sets 4](#_Toc137461619)

[2 Key features of LDS 8 6](#_Toc137461620)

[2.1 Geography 6](#_Toc137461621)

[2.2 Data cleaning and manipulation 6](#_Toc137461622)

[3 Online resources 9](#_Toc137461623)

[3.1 Useful maps and additional information 9](#_Toc137461624)

[3.2 Other useful websites 9](#_Toc137461625)

# 1 Overview

### 1.1 Introduction

All AS and A Level Mathematics for first teaching from September 2017 include the requirement to work with one or more given large data sets (LDS). The criteria from the Department from Education say that:

AS and A Level Mathematics specification must require students to:

* become familiar with one or more specific large data set(s) in advance of the final assessment (the data must be real and sufficiently rich to enable the concepts and skills of data presentation and interpretation in the specification to be explored)
* use technology such as spreadsheets or specialist statistical packages to explore the data set(s)
* interpret real data presented in summary or graphical form
* use data to investigate questions arising in real contexts.

Specifications should require students to explore the data set(s) and associated contexts, during their course of study to enable them to perform tasks that assume familiarity with the contexts, the main features of the data and the ways in which technology can help explore the data. Specifications should also require students to demonstrate the ability to analyse a subset of features of the data using a calculator with standard statistical functions.

The pre-release LDS is primarily a resource for the classroom, to encourage the use of real, LDSs when learning statistics. *Some* questions in the assessment, on the statistics content, will be set in the context of the LDS, in such a way as to provide an advantage to students who have spent time exploring the data. However, this is only a small part of the exam and students will not have access to the LDS, or to a computer, in the assessment. The focus of this resource is very much on teaching and learning.

Note that the first bullet point above includes the phrase “to enable the concepts and skills of data presentation and interpretation in the specification to be explored”. This refers to that particular section of the content document, i.e. section D of the OCR Mathematics B specification. You are welcome to explore the full range of concepts and skills, but the focus of the requirement, and therefore of the assessment, is on data presentation and interpretation.

### 1.2 MEI Large Data Set 8 (for AS in June 2025 and A Level in June 2026)

OCR’s MEI Large Data Set 8 consists of data about boroughs in London together with some comparative data for other areas in the UK. Data for the City of London has been included where it is available. The 32 boroughs together with the City of London make up London. Further data are available through the area profiles on the London Datastore.

The data set includes an information sheet which describes the various terms used. Rather than repeat this information here, please refer to that sheet. The terminology and information on the metadata sheet of the LDS is part of the data set, so students are assumed to be at least familiar with that terminology and information. That’s part of the point of putting it there, i.e. that you then know that these are the words students should understand. *However*, the assessment is not a test of memorisation of the details of the data, only that students have worked with the data enough to have some familiarity with the key features. We will take a more in-depth look at *some* features in section 2.

Please note that further commentary on specific aspects of the LDS can be found within the resources on the OCR website for this qualification, on the London Datastore and in resources in Integral.

###

### 1.3 Why three large data sets for the MEI specification?

The large data sets associated with AS and A Levels in Mathematics should serve two purposes: they are a teaching resource and they provide a context for setting examination questions. Our hope is that teachers will use all three for teaching, but for each cohort of students just one will be the focus of some of the questions in the exam. Each data set will be clearly labelled as to when it is used.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | June 2022 | June 2023 | June 2024 | June 2025 | June 2026 | June 2027 |
| **AS** | 5  | 6  | 7 | 8 | 9 | 10 |
| **A Level** | 4  | 5  | 6 | 7 | 8 | 9 |

So if you teach A Level Maths over two years, then the class you start teaching in September 2024 will see some questions on LDS\_8 in their AS exams in 2025 (if they sit AS) and their A Level exams in 2026, as the following table demonstrates.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Publish** | June 2020  | 2021 | 2022 | 2023 | 2024 |
| **Start teaching** | Sept 2021  | 2022 | 2023 | 2024 | 2025 |
| **AS Exam (if sat)** | June 2022  | 2023 | 2024 | 2025 | 2026 |
| **A Level Exam** | June 2023  | 2024 | 2025 | 2026 | 2027 |
| **Large Data Set** | 5  | 6 | 7 | 8 | 9 |

MEI and OCR have some experience of pre-release data from our Core Maths B qualification. The CIA World Factbook data set that forms the current pre-release for that qualification became the basis for our thinking and development for the reformed AS and A Level (LDS 1). We tried to write different types of questions using that data set, based on A Level content. When doing this, we realised that things in some countries have changed quite a lot during the lifetime of the legacy mathematics specifications so the data set would need to be updated from time to time - we didn’t want students learning about how things used to be in the world 15 years ago if that no longer reflected the current position.

We were aware that some students (and maybe teachers) did not enjoy the statistics in the legacy Mathematics A Levels. We think that this may be because in mathematics the focus has traditionally been on learning statistical techniques without much focus on why you might want to use them. The large data sets provide a context to use the techniques and interpret the results.

The use of large data sets in teaching and examining A Level Mathematics is new – it is an opportunity to make the statistics students learn more similar to the ways they will use statistics in future study and work. We thought it was important to review the data sets used and to make sure they continued to be suitable for examining. This needs a three-year cycle – two years for using the data set in teaching and a year to review and update if necessary. LDS 7 was a refreshment of the data from LDS 1 and LDS 4, Similarly LDS 8 is the refreshment of LDS 2 and LDS 5. LDS 9 may be a refreshment of LDS 6, or a new data set, dependent on the post assessment review of the questions set in the live assessment.

### 1.4 Features of the three large data sets

The data in the CIA World Factbook is grouped by country; we realised that data based on individuals would allow better teaching of distributions. There aren’t many publicly available data sets which contain ungrouped data on individuals. The NHANES data set, from American health surveys, is often used in statistics courses and it contains a wealth of data so we decided to use that as one data set.

Having got data about countries and data about (American) individuals, we thought it would be good to have some England-based data – the London Datastore is a good place to find suitable data and so we ended up with the following three initial data sets which we hope will appeal to students with different interests in terms of other subjects they are taking.

* [LDS\_1 Data about countries](https://www.ocr.org.uk/Images/400305-large-data-set-lds_1-h630-2018-h640-2018-2019-pre-release-data-mei.xlsx)
* [LDS\_2 Data about boroughs of London and the regions of England](https://www.ocr.org.uk/Images/400306-large-data-set-lds_2-h630-2019-h640-2020-pre-release-data-mei.xlsx)
* [LDS\_3 Health data about individuals](https://www.ocr.org.uk/Images/400307-large-data-set-lds_3-h630-2020-h640-2021-pre-release-data-mei.xlsx)
* [LDS\_4 Data about countries](https://www.ocr.org.uk/Images/546218-large-data-set-4-2022-series-.xlsx)
* [LDS\_5 Data about boroughs of London and the regions of England](https://www.ocr.org.uk/Images/589680-large-data-set-5-2023-series-.xlsx)
* [LDS\_6 Heath data about individuals](https://www.ocr.org.uk/Images/618038-large-data-set-6-2024-series-.xlsx)
* [LDS\_7 Data about countries](https://www.ocr.org.uk/Images/654620-large-data-set-7-2025-series-.xlsx)

We wanted to make the process of working with data manageable for teachers, educationally valuable for students and workable for examining. We decided that three data sets – one per cohort – updated on a rotating cycle would do the trick. In the first year of teaching the new specifications, teachers might choose to work with one data set. The next year, they could still use the lessons that had gone well as well as introducing the next data set and so on.

# 2 Key features of LDS 8

### 2.1 Geography

The data include area codes which are used by the ONS when publishing regional information; this will make it easier to use additional data when working with the LDS. The London boroughs are included in the OCR Specification A LDS; you can get census data for the London boroughs from there.

### 2.2 Data cleaning and manipulation

This list includes a few details covering data cleaning issues, and any data manipulation which was done to create a single data set from all the fields of data, along with a couple of possible stumbling blocks to watch out for.

1. Data fields. The Information Sheet in the LDS gives details of where to download the data from.
2. Inner and outer London. For a comparison of two different definitions of inner and outer London see <https://en.wikipedia.org/wiki/Outer_London> . The ONS definition is used in the LDS. The box-and-whisker plots below show median house prices for inner and outer London for 2004. Using filtering in Excel allows the relevant data to be selected easily. You may want students to explore what difference it makes if Newham and Haringey are classed as outer London rather than inner London.



1. Drawing graphs and charts. The LDS is a set of time series. You may want to draw the following types of graphs and charts. In each case you will probably find it easier to copy the data you are working with to another Excel spreadsheet; this will make it easier to select the data you want.
	* graphs for one time series for regions of London or for regions of England
	* a series of box-and-whisker plots to compare years within a time series or to compare regions of London with regions of England for one particular year
	* a scatter diagram using data from two different time series for the same year for either regions of London or for regions of England.

 You may want to compare a graph for the regions of London with the same graph for the regions of England but they should usually be done on different axes as London as a whole is included in the English regions.

1. Mean and Median income of taxpayers. The data are based on a survey; further statistics can be downloaded at <https://www.gov.uk/government/collections/personal-incomes-statistics>
2. Missing data. You might want to draw scatter diagrams between different data fields using Excel. For example, a scatter diagram of GCSE results 2011-12 against median house price 2011 for the London boroughs.

Excel draws the scatter diagram correctly as follows:

 Copying the two data columns into GeoGebra gives the following graph. It may not be obvious that it is not the same but the point at about (600 000, 70) in the Excel graph, representing Westminster does not exist in the GeoGebra graph but the point at about (400 000, 80) in the GeoGebra graph does not exist in the Excel graph or in the LDS. GeoGebra has not interpreted the #N/A symbols correctly. If you try to find the correlation coefficient in Excel, the existence of #N/A in some cells will prevent the automatic correlation function from calculating.



 Different software uses different methods of showing that a data value is missing. #N/A ensures that Excel draws graphs correctly but will prevent use of formulas on the data. You may want to use the filter function in Excel to filter out rows which have #N/A in the fields you are working with before copying into other software or when doing a calculation. Some fields have all or nearly all data so you won’t want to delete all rows which include #N/A anywhere.

1. London and its boroughs. If you find summary statistics for the data set of median house prices of the 33 London boroughs for 2013, the median is 326 475 and the mean is 367 983 (to the nearest £). The median house price for London as a whole is 323 000. The value for the whole of London for any data set cannot be found by averaging the borough values because each borough value represents a data set of different size to the other borough values.

# 3 Online resources

### 3.1 Useful maps and additional information

Please note, these links are offerred as additional informaion which students may find interesting and useful; it is not expected that you would use all of these with students.

**London area profiles**

The London borough profiles from <https://data.london.gov.uk/london-area-profiles/> give a great deal of information about all the London boroughs.

**Regions of England**

<https://ec.europa.eu/eurostat/cache/RCI/#?vis=nuts2.labourmarket&lang=en> Life expectancy, unemployment and more broken down by regions of EU countries. The regions of England in the LDS are NUTS1 regions in Eurostat (see <https://en.wikipedia.org/wiki/NUTS_1_statistical_regions_of_England> and <https://en.wikipedia.org/wiki/NUTS_statistical_regions_of_the_United_Kingdom>). You can zoom in to the Eurostat map to just see the UK if you don’t want to compare with the rest of Europe.

### 3.2 Other useful websites

As mentioned earlier, a classic spreadsheet is not the only way to interact with the LDS. These links are a starting point for exploring other tools.

**Tinkerplots**

TinkerPlots is a simple, but powerful, data visualisation and modelling tool developed for use by schools.

<https://www.tinkerplots.com/>

**CODAP**

Online data visualisation based on TinkerPlots and Fathom – just drag a CSV file in and off you go – free to use and needs no download.

<https://codap.concord.org/>

**Geogebra**

Geogebra is a free dynamic mathematics tool, including graphing, 3D graphing, geometry, CAS and (most importantly) a spreadsheet. The website also hosts a vast collection of materials.

<https://www.geogebra.org/>

**JASP**

This is free statistical software which has been developed with the support of the University of Amsterdam. It is fairly intuitive to use for people who can use spreadsheets, comes with online support materials (in English) and has some features which are not available in either Excel or GeoGebra. There are more features than are needed for A Level but this software is a good starting place if you want to try specialist statistical software. For Windows, MAC or Linux.

<https://jasp-stats.org/>

**R**

Finally, R is a free software environment for statistical computing and graphics. It is readily available on a wide variety of operating systems. The interface and language may take some getting used to, but the flexibility and power rewards the effort.

<https://www.r-project.org/>

We’d like to know your view on the resources we produce. By clicking on ‘Like’ or ‘Dislike’ you can help us to ensure that our resources work for you. When the email template pops up please add additional comments if you wish and then just click ‘Send’. Thank you.

If you do not currently offer this OCR qualification but would like to do so, please complete the Expression of Interest Form which can be found here: [www.ocr.org.uk/expression-of-interest](http://www.ocr.org.uk/expression-of-interest)

**OCR Resources**: *the small print*OCR’s resources are provided to support the delivery of OCR qualifications, but in no way constitute an endorsed teaching method that is required by the Board, and the decision to use them lies with the individual teacher. Whilst every effort is made to ensure the accuracy of the content, OCR cannot be held responsible for any errors or omissions within these resources.

Our documents are updated over time. Whilst every effort is made to check all documents, there may be contradictions between published support and the specification, therefore please use the information on the latest specification at all times. Where changes are made to specifications these will be indicated within the document, there will be a new version number indicated, and a summary of the changes. If you do notice a discrepancy between the specification and a resource please contact us at:

[resources.feedback@ocr.org.uk](file:///%5C%5Cfilestorage%5COCR%5CPD%5CProdSup%5CDesign%5CStudio%5CVisual%20Style%20Guidelines%5C2016_Templates%5Cresources.feedback%40ocr.org.uk).

© OCR 2020 - This resource may be freely copied and distributed, as long as the OCR logo and this message remain intact and OCR is acknowledged as the originator of this work. OCR acknowledges the use of the following content: n/a

Please get in touch if you want to discuss the accessibility of resources we offer to support delivery of our qualifications: resources.feedback@ocr.org.uk