National Unit Specification

General information

Unit title: Data Science (SCQF level 6)

Unit code: J8LW 46

Superclass: RB

Publication date: October 2024

Source: Scottish Qualifications Authority

Version: 01

Unit purpose

The purpose of this unit is to develop learners' knowledge and skills in data science. The unit focuses on the key concepts involved in data science and the main methods of data analysis and provides an opportunity for learners to apply this knowledge in a practical context using large datasets. It is desirable, but not required, that learners possess previous knowledge and experience of data science. Computational and numerical competency are essential.

This unit is particularly relevant to learners with a vocational interest in science, technology, engineering, and mathematics (STEM)or those who intend to progress to higher level learning (in any subject).

The unit covers a variety of topics relating to data science including: the applications of data science, potential data bias, methods of data analysis, including statistics, and how to present data using visualisations. Learners will gain practical skills in the analysis of large datasets using contemporary software.

At the completion of this unit, learners will appreciate the principles of data science, understand the various stages in data analysis, and be able to apply this knowledge to real-world problems using software to identify patterns and trends in data.

National Unit Specification: General information (continued)

Unit title: Data Science (SCQF level 6)

Learners may wish to undertake this unit alongside J890 46 Data Citizenship at SCQF level 6, which explores the less technical, societal aspects of data science. Learners may focus on specific aspects of data science by undertaking specialist units alongside this unit such as J891 46 Machine Learning at SCQF level 6 or J2G8 46 Data Science: Statistics at SCQF level 6.

Outcomes

On successful completion of the unit the learner will be able to:

1. Describe the applications of data science.
2. Explain techniques for analysing a dataset.
3. Analyse data to extract insights.

Credit points and level

1 National Unit credit at Scottish Credit and Qualifications Framework (SCQF) level 6: (6 SCQF credit points at SCQF level 6).

Recommended entry to the unit

Previous knowledge and experience of data science (or, at least, computer science and statistics) is recommended. This could be evidenced by possession of J8LW 45 Data Science at SCQF level 5. It is possible for learners to undertake this unit without previous knowledge or experience of data science. However, proficiency in computing and numeracy is required, which could be evidenced by possession of the Core Skill units in Numeracy and Information and Communication Technology (ICT) at SCQF level 6.

Core Skills

Opportunities to develop aspects of Core Skills are highlighted in the support notes for this unit specification.

There is no automatic certification of Core Skills or Core Skill components in this unit.

National Unit Specification: General information (continued)

Unit title: Data Science (SCQF level 6)

Context for delivery

If this unit is delivered as part of a group award, it is recommended that it should be taught and assessed within the subject area of the group award to which it contributes. For example, if this unit is delivered as part of the National Progression Award in Data Science at SCQF level 6 there is overlap with other units within this award (particularly J890 46 Data Citizenship) and there will be opportunities to contextualise and integrate teaching, learning and assessment across component units. There is particular scope for integration with J2G8 46 Data Science: Statistics at level 6, which would permit learners to gain a deeper appreciation of the statistical techniques involved in data science or with J2GT 46 Data Science: Project to further apply the practical skills gained in this unit.

The Assessment Support Pack (ASP) for this unit provides assessment and marking guidelines that exemplify the national standard for achievement. It is a valid, reliable and practicable assessment. Centres wishing to develop their own assessments should refer to the ASP to ensure a comparable standard. A list of existing ASPs is available to download from SQA's website: Internal Assessment Support Materials (www.sqa.org.uk/internal assessment support materials).

Equality and inclusion

This unit specification has been designed to ensure that there are no unnecessary barriers to learning or assessment. The individual needs of learners should be taken into account when planning learning experiences, selecting assessment methods or considering alternative evidence.

Further advice can be found on our website:
SQA Assessment Arrangements (www.sqa.org.uk/assessmentarrangements).

National Unit Specification: Statement of standards

Unit title: Data Science (SCQF level 6)

Acceptable performance in this unit will be the satisfactory achievement of the standards set out in this part of the unit specification. All sections of the statement of standards are mandatory and cannot be altered without reference to SQA.

Outcome 1

Describe the applications of data science.

Performance criteria

(a) Describe contemporary applications of data science and the types of problem that data science can address.
(b) Explain the steps in solving a problem using data science, including the potential for bias at each stage.
(c) Describe descriptive analytics and predictive analytics.
(d) Explain techniques for keeping data secure.

Outcome 2

Explain techniques in analysing a dataset.

Performance criteria

(a) Describe common data types and data formats including structured and unstructured data.
(b) Explain techniques for data capture, cleaning and manipulation.
(c) Explain the use of descriptive statistics used to summarise a dataset.
(d) Explain techniques for data visualisation and data storytelling, including accessibility considerations.

Outcome 3

Analyse data to extract insights.

Performance criteria

(a) Plan an analysis of a dataset to solve a problem.
(b) Identify potential sources of bias in a dataset.
(c) Tidy, clean and manipulate a dataset.
(d) Perform analyses on the data.
(e) Create accessible data visualisation to extract insights.
(f) Make recommendations based on conclusions and communicate findings.

National Unit Specification: Statement of standards (continued)

Unit title: Data Science (SCQF level 6)

Evidence requirements for this unit

Evidence is required to demonstrate that learners have achieved all outcomes and performance criteria.

Learners must provide **knowledge** and **product** evidence.

The knowledge evidence will relate to outcome 1 and outcome 2. The knowledge evidence may be written or oral or a combination of these. The amount of evidence may be the minimum required to infer competence across both outcomes but sufficient for assessors to make assessment judgements with confidence. The descriptions and explanations must demonstrate an understanding of the principles and techniques defined in the respective outcomes.

Knowledge evidence may be sampled when testing is used. Testing must be carried out under supervised conditions and must be controlled in terms of location and time. Access to reference material is not permitted for outcomes 1 and 2. The sampling frame, on all occasions, must include outcome 1 and outcome 2 (but not every performance criterion within each outcome). Given the conceptual and explanatory nature of these outcomes, constructed response questions should be used.

The product evidence will relate to outcome 3. The product evidence will take the form of analysis of a one or more dataset. The dataset should be unfamiliar and complex in terms of data types and size and should require significant cleaning and transformation. The analysis must include a number of visualisations which must provide useful insights into the data and be accessible for a diverse audience.

The analysis will demonstrate that learners can:

- identify potential sources of bias such as missing data then clean and manipulate datasets
- calculate descriptive statistics including measures of central tendency and variability
- use appropriate data visualisations to tell a story and make recommendations

The summary must include insights to answer an identified problem.

The analysis may be done in lightly controlled, open-book conditions, over an extended period of time, at times and places at the discretion of the learner. The evidence must be produced by the learner, alone, without assistance.

The SCQF level of this unit (level 6) provides additional context on the nature of the required evidence and the associated standards. Appropriate level descriptors should be used when making judgements about the evidence.

When evidence is produced in loosely controlled conditions it must be authenticated. The guide to assessment provides further advice on methods of authentication.

The support notes section of this specification provides specific examples of instruments of assessment that will generate the required evidence.

National Unit Support Notes

Unit title: Data Science (SCQF level 6)

Unit support notes are offered as guidance and are not mandatory.

While the exact time allocated to this unit is at the discretion of the centre, the notional design length is 40 hours.

Guidance on the content and context for this unit

This unit is intended for learners who wish to develop existing knowledge and skills in data science. However, it can be undertaken by new learners so long as they possess well-developed computing and numerical skills and have the capacity for rapid learning.

This unit is one in a series of units, with rising difficulty, that relate to data science. This is the last unit in that series and is the most demanding. There is no requirement to undertake the units in sequence since each unit can be attempted without previous knowledge or experience of the subject. However, learners without previous knowledge or experience of data science or data analysis will face a steep learning curve.

The aim of the unit is to develop skills in the analysis of large datasets using contemporary data analysis tools. It is intended for a wide range of learners, particularly those who are progressing to higher level studies (in any subject), who will benefit from acquiring data skills prior to progression.

It is suggested that the unit covers the following knowledge and skills.

Knowledge	**Skills**
Applications of data science (including artificial intelligence (AI). Ethical issues in Data Science. Methods to clean and transform data. Differences between structured and unstructured data. Data quality including data bias.	Clean a more complex dataset: • Same as level 5 but using automated process rather than manual ones • Transform dataset • Combining datasets (joining datasets using keys columns, and more than one key column)

National Unit Support Notes (continued)

Unit title: Data Science (SCQF level 6)

Knowledge	Skills
Data types and data formats (level 5 data types with the addition of chars, strings, data frames and arrays). Types of data including nominal, ordinal, categorical, discrete and continuous. Data security. Descriptive statistics: • Sum / totals • Averages (mean, median and mode) • Range • Percentiles • Standard deviation Graphs and charts: • Bar chart • Line chart • Pie chart • Histogram • Scatter plot • Box plot • Area chart Descriptive and predictive analytics.	Calculate descriptive statistics: • Measures of central tendency (mean, median and mode) and when it's appropriate to use each • Measures of variability (standard deviation) Select appropriate graphs and charts. Perform an analysis on a dataset. Create accessible data visualisation to extract insights, for example: • Select an appropriate graph / chart: ○ other visualisations not covered in levels 4 and 5, for example bubble plots, time series graphs or waffle charts, histograms, box plots • Data storytelling • Make recommendations based on conclusions and communicate findings

Learners will require access to appropriate software to undertake this unit. A range of software could be used to provide the required functionality, including dedicated data analysis or visualisation software (such as Jupyter Notebook (trademark), Tableau (trademark), or Power BI (trademark), and specialised programming languages (such as Python and R). It is expected that if learners use a generic application software (such as Microsoft Excel (trademark), then some form of automation is used, not manual processing of the data. While not a requirement, learners could be exposed to more than one toolset to appreciate the strengths and limitations of each.

National Unit Support Notes (continued)

Unit title: Data Science (SCQF level 6)

The selection of appropriate data is important for teaching and learning. The datasets used for teaching and learning should be large, varied and include familiar and unfamiliar contexts. It is not appropriate to focus learning on small, familiar datasets. A critical objective of this unit is to demonstrate the size of contemporary datasets and the need for specialist tools to handle them. Familiar data will be easier for learners to understand and analyse but unfamiliar data should also be used to reinforce learning in unfamiliar contexts. It is recommended that learners use real data to improve the authenticity of learning.

There are many sources of authentic data including services such as Kaggle (Kaggle) For formative purposes, artificially generated data may be useful and can be found from sources such as Mockeroo (Mockaroo).

There is a Trello board of datasets available at Data Sets grouped by topic. These are collated by Data Education in Schools (Data Education in Schools) and contributions from educators and learners are welcomed.

The development of learners' technical vocabulary is vital. Terminology should be introduced, in context, throughout the unit. Learners should be encouraged to use the correct technical terms at all times.

The ethical implication of data science should be emphasised throughout this unit. Reference should be made to national and international standards for data ethics, including the UK's Data Ethics Framework (Data Ethics Framework).

This unit has three outcomes. Outcome 1 introduces learners to the principles behind data science and explains the growing importance of the discipline. Outcome 2 is less conceptual and looks at the techniques involved in the data science process. Outcome 3 is practical; it applies this knowledge to the analysis of a large dataset using a data analysis toolkit.

Outcome 1: This outcome provides a conceptual framework for data science. For some learners, this will be their introduction to the terms and concepts in data science so special care should be given to contextualise each new term and concept.

At this level, detailed explanations are required. For example, the descriptions of contemporary applications of data science (performance criterion (a)) should not only describe what data science is used for (in a specific context) but, also, how it is used.

National Unit Support Notes (continued)

Unit title: Data Science (SCQF level 6)

The performance criteria are relatively self-explanatory. Performance criterion (a) seeks to describe the use of data science and the types of problems they solve (including the use of AI). Learners are not required to use predictive analytics (performance criterion (c)) but should appreciate the distinction between descriptive and predictive analytics.

Data security should be explained in context throughout the unit (performance criterion (d)).

Outcome 2: This outcome applies the principles introduced in outcome 1 to the data science life cycle. It is recommended that all learning takes place in the context of a specific data analysis package. For example, learning about data types and data formats (Performance Criterion (a)) would best be done in the context of the data types and formats supported by a particular toolset.

At this level, learners should be introduced to fairly sophisticated techniques for capturing, cleaning and manipulating data (performance criterion (b)). For example, in the context of Microsoft Excel (trademark), learners would be expected to learn how to use Power Query and Excel's data modelling features or the use of Python libraries such as Pandas, MatPlotlib and Seaborn to capture, clean and structure data. Learners may be unaware of the messy nature of most data and the time-consuming nature of data cleaning and transformation, prior to analysis.

At this level, learners are expected to make relatively sophisticated choices with respect to how they communicate their analyses through visualisations (performance criterion (d)). They should be introduced to a wide range of types of visualisations (beyond those listed at level 5) for storytelling.

Outcome 3: This outcome applies the knowledge gained in outcome 1 and outcome 2 to the analysis of a large dataset. The outcome covers the full data analysis life cycle from capture to presentation. Learning should involve the use of large datasets to emphasise the need for automated processing of data rather than manual processing.

Given the size and complexity of the datasets that learners will use, the importance of analysis (performance criterion (a)) should be emphasised.

Work should be undertaken to understand the dataset through exploratory data analysis (performance criterion (a)). It is preferable that data used in analysis requires learners to manipulate datasets (performance criterion (c)).

Learners are required to identify and create appropriate means of communicating their findings (performance criterion (e and f)).

National Unit Support Notes (continued)

Unit title: Data Science (SCQF level 6)

Guidance on approaches to delivery of this unit

The following distribution of time is suggested.

Outcome 1: 12 hours
Outcome 2: 12 hours
Outcome 3: 16 hours

This unit is a mixture of theory and practice. Outcome 1 and outcome 2 relate to theory and outcome 3 relates to practice.

It is recommended that the outcomes are taught in sequence. Outcome 1 provides a broad introduction to the subject, outcome 2 introduces analytical methods, and outcome 3 applies this knowledge to the analysis of a dataset.

However, there is scope to combine outcome 2 and outcome 3 so that learners are introduced to techniques in outcome 2 and immediately practice those techniques, using appropriate software, in outcome 3. For example, once techniques for data cleaning and transformation are introduced in outcome 2 (performance criterion (b)), they can be contextualised and practiced using a specific software tool in outcome 3.

It is recommended that a problem-solving approach is taken to teaching and learning. Learners should develop their knowledge and skills in the context of different problems, with varying complexity, relating to a variety of datasets. It is important that the datasets used are large and require significant analysis, cleaning, structuring and manipulating. Large datasets may be split so that learners can demonstrate joining techniques. It may be helpful to learners to expose them to examples of data analysis and data visualisation by using services such as Our World in Data (Our World in Data). There is a Trello board of suitable data visualisation examples (Data Visualisations) and 'bad graphs' (Bad Graphs) grouped by visualisation type and complexity.

There are many sources of engaging content about data science that will aid the delivery of outcome 1. For example, there are many case studies relating to the applications of data science, describing how it can be used in a wide range of fields. There is a Trello board of examples and resources for this outcome (Data Science Resources and links).

Outcome 2 may be learners first exposure to data science techniques and will require care in the way that it is taught. Learning can be enlivened through the use of videos and real-world examples.

National Unit Support Notes (continued)

Unit title: Data Science (SCQF level 6)

Outcome 3 may be learners first experience of analytical software. The learning curve will be significantly reduced if this environment is already familiar to learners (such as Excel (trademark), or Python) rather than an entirely new product. There are Python and Excel lessons designed for this outcome available for educators and learners (Learn Data Science). The lessons come with Powerpoint presentations, lesson plans, activities in a choice of format (either Excel workbook or interactive Python notebook).

Guidance on approaches to assessment of this unit

Evidence can be generated using different types of assessment.

The following are suggestions only. There may be other methods that would be more suitable to learners and the type of learner assessment activities will vary depending on the resources available.

Centres are reminded that prior verification of centre-devised assessments would help to ensure that the national standard is being met. Where learners experience a range of assessment methods, this helps them to develop different skills that should be transferable to work or further and higher education.

Summative assessment may be carried out at any time. However, when testing is used (see evidence requirements) it is recommended that this is carried out towards the end of the unit (but with sufficient time for remediation and re-assessment). When continuous assessment is used, this could commence early in the unit and be carried out throughout the life of the unit.

A wide range of instruments of assessment could be used to satisfy the evidence requirements. A traditional approach to assessment could involve the use of an extended response test for knowledge evidence and a practical assignment for product evidence.

The extended response test could comprise a sight-unseen question paper, sampling from the knowledge domain (outcome 1 and outcome 2). The questions would relate directly to the performance criteria but may combine two or more performance criteria. A rubric would assign marks to each response. An appropriate pass mark would be set (for example, 50%).

The assessment for outcome 3 could be a practical assignment, which requires learners to source, clean, transform and analyse a large, relatively complex dataset, including appropriate visualisations, and from that analysis draw conclusions and make recommendations.

National Unit Support Notes (continued)

Unit title: Data Science (SCQF level 6)

More contemporary approaches to assessment include the use of a web log or the creation of a portfolio. The web log would record learning over the life of the unit. Practical work could be recorded on the blog in a variety of ways (for example, specific posts could link to completed analyses). The completed blog would have to satisfy all performance criteria. The blog would be assessed on a pass / fail basis using a checklist. Alternatively, a portfolio could be used as a repository for the descriptions and explanations required in outcome 1 and outcome 2, and the output from learners' practical work in outcome 3. The completed portfolio would have to satisfy all performance criteria. The portfolio would be assessed on a pass / fail basis using a checklist.

There are opportunities to carry out formative assessment at various stages in the unit. For example, formative assessment could be carried out on the completion of each outcome to ensure that learners have grasped the knowledge contained within it. This would provide assessors with an opportunity to diagnose misconceptions and intervene to remedy them before progressing to the next outcome.

Opportunities for e-assessment

E-assessment may be appropriate for some assessments in this unit. By e-assessment we mean assessment which is supported by Information and Communication Technology (ICT), such as e-testing or the use of e-portfolios or social software. Centres which wish to use e-assessment must ensure that the national standard is applied to all learner evidence and that conditions of assessment as specified in the evidence requirements are met, regardless of the mode of gathering evidence. The most up-to-date guidance on the use of e-assessment to support SQA's qualifications is available at SQA e-Assessment. (www.sqa.org.uk/Guide to best practice.pdf).

Opportunities for developing Core and other essential skills

The unit is particularly well suited to developing the Core Skills of Numeracy and Information and Communication Technology (ICT). ICT skills will be used throughout the unit, particularly outcome 2 and outcome 3. Numeracy skills will be developed in outcome 2, when learners are introduced to descriptive statistics, and outcome 3, when learners are introduced to visualisations.

The computational thinking skills of abstraction and automation will be developed in this unit when learners create models (abstraction) and perform analyses (automation) using software tools.

Employability skills will be developed when learners gain skills in the use of software to analyse data. For example, data analysis skills are valued by employers.

History of changes to unit

Version	Description of change	Date

Unit template: February 2024

General information for learners

Unit title: Data Science (SCQF level 6)

This section will help you decide whether this is the unit for you by explaining what the unit is about, what you should know or be able to do before you start, what you will need to do during the unit and opportunities for further learning and employment.

Data science is growing field which offers excellent job opportunities to keen, motivated students. This unit provides an opportunity to learn how data is sourced and how data science tools and techniques are used to manipulate and analyse the data. You do not require any previous knowledge or experience of data science. You will also learn the skills, used by data scientists, in preparing, analysing and presenting information.

On completion of this unit, learners will have a good grounding in data science and the importance of data in the world today as well as being able to competently manipulate and interpret datasets.

In outcome 1 you will learn about data science and some of the contemporary uses it is put to in real life. You will learn about the different types of analytics and ethical issues in data science like bias.

You will learn about how data is crucial to recent advancements in ‘AI and Generative AI’.

In outcome 2 you will learn about common data types and formats along with how data is classified. You will learn about automated techniques used to capture, clean and transform a dataset and the types of tools used data science.

You will also learn how statistics can be used to describe a dataset and techniques for data visualisation and communication.

Outcome 3 is a practical outcome where you will put the theory you learned in outcomes 1 and 2 into practise. You will be given datasets to work on throughout this outcome. You will be expected to plan to solve a problem using data before cleaning, transforming and then analysing each dataset. Analysis may include restructuring and joining datasets. Once you have found the information you need you will present it in a way that clearly communicates insights.

You can be assessed in a variety of ways such as writing answers to questions, making a poster or presentation, speaking to answer questions, making a short video or audio recording or working in a group with other learners.