National Unit Specification

General information

Unit title: Data Science (SCQF level 6)

Unit code: J2G2 46

Superclass: RB

Publication date: March 2020

Source: Scottish Qualifications Authority

Version: 03

Unit purpose

The purpose of this unit is to develop learners' knowledge and skills in data analysis. 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 is essential.

This is a **non-specialist** unit intended for all learners. It is particularly relevant to learners with a vocational interest in 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, data ethics, methods of data analysis, and how to present data using dashboards and 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 trends in data and make predictions about the future.

Learners may wish to undertake this unit alongside J2HN 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 J2G6 46 *Machine Learning* at SCQF level 6 or J2G8 46 *Data Science: Statistics* at SCQF level 6.

National Unit Specification: General information (cont)

Unit title: Data Science (SCQF level 6)

Outcomes

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

1 Explain the principles of data science.
2 Explain data science techniques.
3 Analyse a dataset to make predictions.

Credit points and level

1 National Unit credit at 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 J2G2 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

Achievement of this Unit gives automatic certification of the following:

Complete Core Skill	Numeracy at SCQF level 6 Information and Communication Technology at SCQF level 5
Core Skill component	Critical Thinking at SCQF level 5

Opportunities to develop aspects of Core Skills are highlighted in the Support Notes for this Unit specification.

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 J2HN 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.

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 **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.

Where evidence for outcomes is assessed on a sample basis, the whole of the content listed in the knowledge and/or skills section must be taught and available for assessment. Learners should not know in advance the items on which they will be assessed and different items should be sampled on each assessment occasion.

Outcome 1

Explain the principles of data science.

Performance criteria

(a) Explain the relationship between artificial intelligence, machine learning, big data and data science.
(b) Explain the technological, economic and societal reasons for the development and growth of data science.
(c) Describe contemporary applications of data science and the types of problem that data science can address.
(d) Explain the data science life cycle and the significance of domain expertise.
(e) Explain descriptive analytics and predictive analytics.
(f) Explain the principle of open data and sources of open data.
(g) Explain data ethics, including data bias, with reference to national and international standards and frameworks.

Outcome 2

Explain data science techniques.

Performance criteria

(a) Describe common data types and data formats including structured and unstructured data.
(b) Explain techniques for data capture, cleaning and transformation including data modelling.
(c) Explain data management and data security techniques.
(d) Explain statistical techniques involved in data science.
(e) Explain techniques for data visualisation, data dashboards and data storytelling.

National Unit Specification: Statement of standards (cont)

Unit title: Data Science (SCQF level 6)

Outcome 3

Analyse a dataset to make predictions.

Performance criteria

(a) Define the required analyses and data models.
(b) Create a relational data model from external sources of data.
(c) Perform data transformation to complete, correct and structure data.
(d) Perform descriptive and predictive analyses on the data.
(e) Create data visualisations and data dashboards to provide insights.
(f) Identify potential sources of bias in the analysis.

Evidence requirements for this unit

Learners will need to provide evidence to demonstrate the performance criteria across all outcomes. The evidence requirements for this unit will take **two** forms.

1 Knowledge evidence.
2 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.

The 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. The sampling frame, on all occasions, must include Outcome 1 and Outcome 2 (but not every performance criterion within each outcome). The sampling frame must always include data ethics (Outcome 1, Performance Criterion (g)). Given the conceptual and explanatory nature of these outcomes, testing should be restricted to extended response questions.

The **product evidence** will relate to Outcome 3. The product evidence will take the form of a completed analysis of a dataset. The data model will be created by the learner, captured externally from at least two sources, and must comprise at least 10,000 multi-variate records (rows), some of which will require cleansing. The data must be real data, captured from authentic sources. The derived data model must be relational. The data model must be capable of being used for forecasting or predicting. The analysis must include at least one dashboard and a number of visualisations, which must provide useful insights into the dataset, including the ability to forecast or predict. The dashboard must be interactive and provide sophisticated insights into the dataset, providing a range of dynamic data views. It must be appropriately presented and be easy to use.

The analysis may be done in lightly controlled 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.

National Unit Specification: Statement of standards (cont)

Unit title: Data Science (SCQF level 6)

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.

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 software (such as Jupyter Notebook™, Tableau™ or Power BI™), generic application software (such as Microsoft Excel™) and specialised programming languages (such as Python and R). It is recommended that learners are exposed to more than one toolset to appreciate the strengths and limitations of each.

The selection of appropriate data is important for teaching and learning. The datasets used for teaching and learning should be large and 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 (**https://www.kaggle.com/datasets**) and data.world (**https://data.world/**). For formative purposes, artificially generated data may be useful and can be found from sources such as Mockeroo (**https://mockaroo.com/**).

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.

National Unit Support Notes (cont)

Unit title: Data Science (SCQF level 6)

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 (**https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework**). Causes of deliberate and accidental data bias should be introduced in context, with particular reference to the problem of historical data bias. Other forms of bias (such as algorithmic bias) should also be introduced.

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.

Please note that the following guidance does not seek to explain each performance criterion. This section seeks to clarify the statement of standards where it is potentially ambiguous. It also focuses on non-apparent teaching and learning issues that may be over-looked, or not emphasised, during delivery. As such, it is not representative of the actual time spent teaching or learning specific competences or the relative importance of each competence.

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 (c)) should not only describe what data science is used for (in a specific context) but, also, how it is used.

The performance criteria are relatively self-explanatory. Performance Criterion (a) seeks to place data science in the context of artificial intelligence, machine learning and big data. The relationship between data analysis and machine learning should receive special attention. Learners are not required to use predictive analytics (Performance Criterion (e)) but should appreciate the distinction between descriptive analytics and predictive analytics.

Data ethics should be explained in context throughout the unit. However, this outcome makes explicit reference to ethics and bias (Performance Criterion (g)) so particular emphasis should be placed on these topics in this outcome.

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, transforming and modelling data (Performance Criterion (b)). For example, in the context of Microsoft Excel™, learners would be expected to learn how to use Power Query and Excel's data modelling features 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.

National Unit Support Notes (cont)

Unit title: Data Science (SCQF level 6)

Given time constraints, the treatment of data management and data security (Performance Criterion (c)) will be light but should be sufficient to introduce learners to the most important considerations relating to data management and data security from a data science perspective.

At this level, learners are expected to make relatively sophisticated choices with respect to how they present their analyses through visualisations and dashboards (Performance Criterion (e)). They should be introduced to a wide range of types of visualisations (such as time series, comparison and geo-spacial visualisations) and different ways of permitting user interaction through dashboards and other techniques 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, comprising at least 10,000 records.

Given the size and complexity of the datasets that learners will use, the importance of data analysis and design (Performance Criterion (a)) should be emphasised.

The predictive analyses required (Performance Criterion (d)) may be done through the identification of trends using descriptive analytics.

Learners are required to identify potential sources of bias in their analysis (Performance Criterion (f)). This is most likely to come from the data sources but all types of bias should be considered.

Guidance on approaches to delivery of this unit

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 (5,000–50,000 records) and require significant cleaning, structuring and analysing. 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 (**https://ourworldindata.org/**).

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.

National Unit Support Notes (cont)

Unit title: Data Science (SCQF level 6)

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.

Outcome 3 may be learners first experience of analytical software. The learning curve will be significantly reduced if this software is already familiar to learners (such as Excel™) rather than an entirely new product.

A suggested distribution of time is:

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

If Outcome 2 and Outcome 3 are delivered holistically, then the combined time available to learn and practice data analysis methods would be 28 hours.

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.

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. Specimen questions could include:

1 Give **two** sociological reasons for the growth of data science.
2 Explain how machine learning improves data analysis.
3 What is the purpose of data cleaning? Illustrate your answer with at least **two** different examples of data cleaning.
4 What is meant by data ethics? Explain how historical data can introduce bias into a study.
5 Explain the role of subject matter experts with reference to the data science life cycle.

Note that Question 5 is integrative. A rubric would assign marks to each response. An appropriate pass mark would be set (for example, 50%).

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 **www.sqa.org.uk/e-assessment**.

National Unit Support Notes (cont)

Unit title: Data Science (SCQF level 6)

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.

This Unit has the Core Skill of Information and Communication Technology SCQF level 5 embedded in this unit. When a learner achieves the unit, their Core Skills profile will also be updated to include this Core Skill.

This Unit has the Core Skill of Numeracy SCQF level 6 embedded in this unit. When a learner achieves the unit, their Core Skills profile will also be updated to include this Core Skill.

The Critical Thinking component of Problem Solving at SCQF level 5 is embedded in this unit. When a learner achieves the unit, their Core Skills profile will also be updated to include this component.

History of changes to unit

Version	Description of change	Date
03	Clarification in Evidence Requirements section; term ‘data item’ replaced with ‘record’.	04/03/20
02	Core Skill of Information and Communication Technology SCQF level 5 embedded. Core Skill of Numeracy SCQF level 6 embedded in this unit. The Critical Thinking component of Problem Solving at SCQF level 5 is embedded in this unit.	16/08/19

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 a new field within computer science that is becoming increasingly important. It is predicted that there will be a shortage of data scientists in the future.

Data science relates to the analysis of large amounts of data to find patterns and trends so that predictions can be made about the future. Data science is used in a large, and growing, number of fields including astronomy, healthcare and sports science.

Knowledge and skills in data science will be useful to you in various ways. There is a growing number of degree courses and jobs in data science. No matter what job you do, it is likely to involve some aspects of data science. And possessing “analytical skills” is a good life-skill for you. For example, if you plan to progress to university, to do any degree, the ability to analyse large amounts of data, using software such as Microsoft Excel™, is a very useful skill.

Ideally, you should already possess some knowledge of data science before undertaking this unit. But it is possible, with a lot of hard work, to attempt this unit without previous knowledge of the subject.

There are three parts to this unit.

1 Principles of data science.
2 Techniques of data science.
3 Applying data science.

The principles of data science includes data ethics, which is becoming very important as data science makes more decisions for people. You will also learn about how data science relates to artificial intelligence, machine learning and big data. The techniques of data science shows you how to use data science to find patterns and trends in data. And the last part of the unit provides you with practical skills in carrying out an analysis on a large dataset. This will give you the opportunity to learn how to use software to actually carry out data analysis and present your findings using visualisations and data dashboards. You will use software such as Excel™ or Tableau™ or Junyper Notebooks™ to carry out analyses.

The assessment of this unit might involve a test of your knowledge and a practical assignment. For example, you might be asked to explain how data science can be used in healthcare. The practical assignment will involve you in conducting a large-scale data analysis using real data to make predictions about the future.

When you complete this unit you could progress to more advanced studies in data science or simply use your knowledge and skills in another subject area.

General information for learners (cont)

Unit title: Data Science (SCQF level 6)

This Unit has the Core Skill of Information and Communication Technology SCQF level 5 embedded in this unit. When a learner achieves the unit, their Core Skills profile will also be updated to include this Core Skill.

This Unit has the Core Skill of Numeracy SCQF level 6 embedded in this unit. When a learner achieves the unit, their Core Skills profile will also be updated to include this Core Skill.

The Critical Thinking component of Problem Solving at SCQF level 5 is embedded in this unit. When a learner achieves the unit, their Core Skills profile will also be updated to include this component.