National Unit Specification

General information

Unit title: Data Science (SCQF level 4)

Unit code: J2G2 44

Superclass: RB

Publication date: March 2020

Source: Scottish Qualifications Authority

Version: 03

Unit purpose

The purpose of this unit is to introduce learners to the basics of data science. The unit focuses on the **fundamentals** of data science including what it is, how it is used, and how to apply it to small datasets. No previous knowledge or experience of data science is required. However, computational and numerical competency is presumed.

This is a **non-specialist** unit, suitable for all learners. The unit introduces the basic ideas behind data science, what it is used for, including how it might be misused, basic skills in analysing small datasets, and presenting information in a variety of simple visual formats. Statistical methods are introduced in context. The unit will permit learners to gain a familiarity with this emerging field, and improve their appreciation of the growing importance of data science.

At the completion of this unit, learners will understand the basics of data science and the importance of data in the world today, and be able to manipulate and interpret small datasets. Learners may wish to undertake this unit alongside J2HN 44 *Data Citizenship* at SCQF level 4 or progress to more advanced units in this field such as J2G2 45 *Data Science* at SCQF level 5, which will develop the knowledge and skills gained by undertaking this unit.

Outcomes

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

1 Describe data science.
2 Describe simple ways of analysing data.
3 Analyse a small dataset to identify patterns.

National Unit Specification: General information (cont)

Unit title: Data Science (SCQF level 4)

Credit points and level

1 National Unit credit at SCQF level 4: (6 SCQF credit points at SCQF level 4)

Recommended entry to the unit

No previous knowledge or experience of data science is required. However, basic competency in computing and numeracy is required. This may be evidenced by possession of the Core Skills units in *Information and Communication Technology (ICT)* and *Numeracy* at SCQF level 4.

Core Skills

Achievement of this Unit gives automatic certification of the following Core Skills component:

Core Skill component	Accessing Information at SCQF level 4 Using Graphical Information at SCQF level 4 Critical Thinking at SCQF level 3

There are also opportunities to develop aspects of Core Skills which are highlighted in the Support Notes of 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 4 there is overlap with other units within this award (particularly J2HN 44 *Data Citizenship*) and there will be opportunities to contextualise and integrate teaching, learning and assessment across component units.

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 4)

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

Describe data science.

Performance criteria

(a) State the reasons for the development of data science.
(b) Describe contemporary applications of data science.
(c) Describe the steps in solving a problem using data science.
(d) Identify sources of bias in data science including historical bias.

Outcome 2

Describe simple ways of analysing data.

Performance criteria

(a) Describe common data types and data formats.
(b) Describe structured and unstructured data.
(c) Describe simple methods of cleaning and transforming data.
(d) Describe basic descriptive statistics used to summarise a dataset.
(e) Describe simple data visualisations.

Outcome 3

Analyse a small dataset to identify patterns.

Performance criteria

(a) Perform simple data cleaning and structuring.
(b) Perform basic analyses including sort, filter, group and summarise.
(c) Visualise the data to provide basic insights.
(d) Create a simple report to communicate insights.

National Unit Specification: Statement of standards (cont)

Unit title: Data Science (SCQF level 4)

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. The identifications, statements and descriptions may be straightforward but examples should be provided where appropriate. For Outcome 2, the descriptive statistics may be limited to basic analyses but must include measures of central tendency (minimally mean, median and mode) and measures of dispersion (including range).

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 Outcome 2, Performance Criterion (d).

The **product evidence** will relate to Outcome 3. The product evidence will take the form of a completed analysis of a small dataset. The dataset will be supplied to the learner and must comprise at least 500 records (rows), some of which will require cleaning and structuring. The dataset must be cleaned, structured, sorted, filtered, grouped and summarised. The analysis must include at least one visualisation, which must illustrate patterns in the data.

The evidence must be produced by the learner with limited, or no, assistance. 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 SCQF level of this unit (level 4) 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 4)

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 beginners in data science. It provides a basic introduction to the subject for a wide range of learners. No previous knowledge of computer science, data science or statistics is assumed.

This unit is one in a series of units, with rising difficulty, that relate to data science. This is the first unit in that series and is the most accessible to learners. 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, this unit serves as a gentle introduction to the subject and will serve as a foundation for progressing to higher level units.

The aim of the unit is to show learners what data science is, how it can be used, and how to perform simple analyses on small datasets using contemporary software.

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 Tableau™ or Power BI™), generic application software (such as Microsoft Excel™) and specialised programming languages (such as Python and R). It is recommended that, at this level, familiar software is used such as Microsoft Excel™, which provides all of the required functionality (older versions may require add-ins).

The selection of appropriate data is important for teaching and learning. The datasets used should be (relatively) 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 important. 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 4)

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.

The unit comprises three outcomes. Outcome 1 and Outcome 2 are theoretical, and Outcome 3 is practical.

Outcome 1: This outcome is a basic introduction to the field of data science. It should be assumed that this is the learner's first exposure to the subject. As such, the descriptions should be simple and only cover the fundamentals. For example, the reasons for the development of data science (Performance Criterion (a)) should be high level and broad, rather than narrow and deep. When introducing the applications of data science (Performance Criterion (b)), it is important to illustrate the breadth of application, which spans astronomy, business, cosmology, crime, education, healthcare and science. Detailed descriptions of the application of data science in these fields are not required.

Performance Criterion (c) (the steps involved in data science) should avoid technicalities. A simple: capture, clean, structure, analyse, visualise and report process description is sufficient. There is scope to introduce common sources of data.

The treatment of bias in data science (Performance Criterion (d)) is a complex topic for this level and will require careful introduction. It is sufficient to explain that bias in data will result in bias in decision making. More complex sources of bias (such as algorithmic bias) are not appropriate for this level.

Outcome 2: This outcome relates to simple ways of analysing data. Most of the performance criteria are self-evident. The key aspect of this outcome is the level of treatment, which should be basic (for all performance criteria).

It is recommended that delivery of this theoretical outcome is done in the context of an actual software product. For example, the description of data types and formats (Performance Criterion (a)) could be related to the data types and formats supported by Excel™.

When discussing methods of data cleaning and transformation (Performance Criterion (c)), the reasons for *needing* to clean and transform data should be emphasised, as should the time-consuming nature of data cleaning, which often constitutes the longest, and most labour intensive, part of the data analysis process. At this level, the methods of data cleaning and transformation introduced to learners should be simple. For example, cleaning may be limited to removing empty records, correcting obvious errors and fixing an incorrect data type. Transformations might be limited to some (not much) restructuring of the data in preparation for its subsequent analysis. There is an opportunity to introduce the concepts of grouping and categorising (coding) data.

The basic descriptive statistics (Performance Criterion (d)) should be limited to simple summary statistics and common measures of central tendency and dispersion such as sum, mean, median, mode and range. Learners are required to describe these statistics, including worked examples.

National Unit Support Notes (cont)

Unit title: Data Science (SCQF level 4)

It is anticipated that a significant part of this outcome will be spent describing types of data visualisation (Performance Criterion (e)). It is recommended that learners explore how visualisations are used to represent complex datasets such as population growth and historical trends in poverty. Websites such as Our World in Data (**https://ourworldindata.org/**) are good sources for this information. Learners should appreciate how different types of data are best visualised using different infographics.

Outcome 3: This outcome applies the knowledge gained in Outcome 1 and Outcome 2. Learners are required to analyse a small dataset. Datasets should be around 500 records.

Since this will be learners first exposure to analytical software (or their first exposure to the analytical features of familiar software such as Excel™) some time will be required to gain basic familiarity with the software and its analytical features. The terminology of data analytics (“clean”, “transform”, etc.) will require careful introduction.

Learners are not required to capture data. Data will be supplied to learners. To motivate learners, it is recommended that the data relates to topics of interest to learners.

At this level, it is sufficient to frame learning around small, familiar datasets that learners use to practice basic analytical techniques using appropriate software. A significant part of this outcome could be spent exploring data visualisations (Performance Criterion (c)), describing best practice in their selection and construction for different types of data.

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 field, Outcome 2 introduces basic analytical methods, and Outcome 3 applies this knowledge to the analysis of a small dataset.

However, there is scope to combine Outcome 2 and Outcome 3 so that learners are introduced to methods in Outcome 2 and immediately practice those methods, using appropriate software, in Outcome 3. For example, once basic descriptive statistics are described in Outcome 2 (Performance Criterion (d)), learners can use software to calculate these statistics for a variety of small datasets (Outcome 3, Performance Criterion (b)).

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. For example, learners could be supplied with a (fictitious) dataset comprising 300 examination scores and attempt to answer specific questions relating to that dataset (such as the effects of changing pass marks or gender differences in examination scores).

Learners will require access to computing resources, including software capable of analysing data.

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 4)

Outcome 2 provides learners first exposure to data analysis, 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 is likely to be learners first experience of applying data analysis. 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: 10 hours
- Outcome 2: 15 hours
- Outcome 3: 15 hours

If Outcome 2 and Outcome 3 are delivered holistically, then the combined time available to learn and practice data analysis methods would be 30 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 a selected response test for knowledge evidence and a practical exercise for product evidence. The selected response test could comprise a multiple choice test of learners' knowledge of Outcome 1 and Outcome 2. The questions would relate to the identifications, statement and descriptions defined in the performance criteria. The test would sample from the knowledge domain (Outcome 1 and Outcome 2). Note that every test must include questions about descriptive statistics. An appropriate pass mark would be set. The practical exercise would lead learners through the steps required to analyse a small, supplied dataset and produce a simple data visualisation and report, based on the data. A checklist could be used to assess the completed analysis.

More contemporary approaches to assessment include the creation of a web log or portfolio. The web log (blog) would record learning over the life of the unit. The blog would record, on a daily or weekly basis, the learning and activities undertaken by each learner. Practical work could be captured in the blog by linking specific post(s) to examples of analyses carried out by the learner. The completed blog would have to satisfy all performance criteria. Alternatively, a portfolio could be used as a repository for the identifications, statements and descriptions 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.

National Unit Support Notes (cont)

Unit title: Data Science (SCQF level 4)

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**.

Opportunities for developing Core and other essential skills

There are opportunities in this unit to develop Core Skills, computational thinking skills and employability 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 3. Numeracy skills will be developed in Outcome 2, when learners are introduced to descriptive statistics and 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 in Outcome 3, when learners gain skills in the use of software to analyse data. For example, skills in using spreadsheets are valued by employers.

The Accessing Information component of Information and Communication Technology at SCQF level 4 is embedded in this unit, the Using Graphical Information component of Numeracy at SCQF level 4 is embedded in this unit and the Critical Thinking component of Problem Solving at SCQF level 3 is embedded in this unit.
When a learner achieves these units, their Core Skills profile will also be updated to include these components.

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 Skills Components Accessing Information at SCQF level 4 Using Graphical Information at SCQF level 4, Critical Thinking at SCQF level 3 embedded.	16/08/19

General information for learners

Unit title: Data Science (SCQF level 4)

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.

This unit will provide a basic introduction to data science. No previous knowledge or experience of the subject is required.

Data science is becoming very important. Data science is the process of exploring large amounts of data to identify patterns and trends and make predictions. For example, data science is used to discover cancers, find new planets and predict crime.

Data science is a growing field. It is expected that there will be many jobs in this area in the coming years. Every person, no matter their job, will require some knowledge of data science. It is also a useful skill for your future learning, no matter what subject interests you.

There are three parts to this unit.

1 Introduction to data science.
2 Introduction to data analysis.
3 Using software to analyse data.

The introduction to data science explores how it is used in lots of areas such as astronomy, crime fighting and healthcare. You will also understand how data science can go wrong because of bias.

The introduction to data analysis looks at simple ways of exploring data to find patterns and trends. You will be introduced to some of the terminology used in the field such as “data cleaning” and “data structuring”. This part of the unit will also introduce some basic statistics that are commonly used in data science.

The final part of the unit looks at how to analyse data. You will use software to carry out an analysis of a small dataset to find patterns and trends in the data. This part of the unit will give you practical skills in using data analysis software. This will include learning how to present data using graphs and charts and other visualisations.

You are likely to learn about data science in a variety of ways. You might read case studies and watch videos about how it is used, and explore data about music or health or sport using software such as Microsoft Excel™.

The assessment of this unit might involve a test of your knowledge and a practical exercise. Most of your time will be spent learning about data science. Assessment will not take much time.

The Accessing Information component of Information and Communication Technology at SCQF level 4 is embedded in this unit, the Using Graphical Information component of Numeracy at SCQF level 4 is embedded in this unit and the Critical Thinking component of Problem Solving at SCQF level 3 is embedded in this unit.
When a learner achieves these units, their Core Skills profile will also be updated to include these components.

When you complete this unit you could learn more about data science by doing advanced units in this subject area such as *Data Science* at SCQF level 5.