National Unit Specification

General information

Unit title: Data Science (SCQF level 5)

Unit code: J2G2 45

Superclass:	RB
Publication date:	March 2020
Source:	Scottish Qualifications Authority
Version:	03

Unit purpose

The purpose of this unit is to introduce learners to data science in today's world. The unit focuses on the tools and techniques involved in data science, the main methods of data analysis, and provides an opportunity for learners to apply this knowledge in a practical context. 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, particularly those who require data skills prior to commencing university or employment. The unit covers a variety of topics relating to data science including: the reasons for the emergence of data science as a distinct discipline, the uses and misuses of data and data science, the data science life cycle and common methods of data analysis. Learners will also gain practical skills in using software to identify patterns and trends in data.

At the completion of this unit, learners will appreciate the basic principles of data science and be able to apply this knowledge to solve routine problems using data analysis software. Learners may wish to undertake this unit alongside J2HN 45 *Data Citizenship* at SCQF level 5 or progress to more advanced units in this field such as J2G2 46 *Data Science* at SCQF level 6, which explores data science in the context of larger datasets. Learners may focus on particular aspects of data science by undertaking specialist units alongside this unit, such as J2G6 45 *Machine Learning* at SCQF level 5 or J2G8 45 *Data Science: Statistics* at SCQF level 5.

Unit title: Data Science (SCQF level 5)

Outcomes

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

1 Describe the tools and techniques of data science.
2 Describe methods of routine data analysis.
3 Analyse a dataset to identify patterns and trends.

Credit points and level

1 National Unit credit at SCQF level 5: (6 SCQF credit points at SCQF level 5)

Recommended entry to the unit

No previous knowledge or experience of data science is required. However, 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 5.

Core Skills

Achievement of this Unit gives automatic certification of the following:

Complete Core Skill	Numeracy at SCQF level 5 Information and Communication Technology at SCQF level 4

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

Core Skill component	Critical Thinking at SCQF level 4

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 5 there is overlap with other units within this award (particularly J2HN 45 *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 45 *Data Science: Statistics* at SCQF level 5, which would permit learners to gain a deeper appreciation of the statistical techniques involved in data science.

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 (**http://www.sqa.org.uk/sqa/46233.2769.html**).

National Unit Specification: General information (cont)

Unit title: Data Science (SCQF level 5)

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 5)

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 the tools and techniques of data science.

Performance criteria

(a) Describe the reasons for the development and growth of data science.
(b) Describe contemporary applications of data science.
(c) Describe the data science life cycle including the potential for bias at each stage.
(d) Describe the tools that can be used at each stage in the life cycle.
(e) Identify sources of public and private datasets.
(f) Describe the role of domain knowledge and subject matter experts in data science.

Outcome 2

Describe methods of routine data analysis.

Performance criteria

(a) Describe common data types and data formats.
(b) Describe the composition of a structured dataset.
(c) Describe methods of cleaning and transforming data.
(d) Describe methods of securing and managing data.
(e) Describe descriptive statistics used to summarise a dataset including measures of central tendency and dispersion.
(f) Describe the selection of data visualisations to illustrate different types of data.

Outcome 3

Analyse a dataset to identify patterns and trends.

Performance criteria

(a) Define the required analyses.
(b) Capture data from an external source.
(c) Perform routine data cleaning and structuring.
(d) Perform analyses including query, sort, filter, consolidate, group and summarise.
(e) Visualise the data to provide insights.
(f) Create an interactive data dashboard to identify patterns and trends.

Unit title: Data Science (SCQF level 5)

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 descriptions may be straightforward but examples should be provided where appropriate. For Outcome 2, the descriptive statistics may be limited to routine statistics but must include measure of central tendency (including mean, median and mode) and measures of dispersion (including range and interquartile 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 (e).

The **product evidence** will relate to Outcome 3. The product evidence will take the form of a completed analysis of a dataset. The dataset will be created (captured) by the learner, sourced externally, and must comprise at least 5,000 records (rows), some of which will require cleaning. The dataset must be cleaned, structured, sorted, filtered, grouped and summarised. The analysis must include at least one dashboard and at least three visualisations. The dashboard must be interactive and provide useful insights into the dataset, providing a range of dynamic data views, and must be presented attractively and be easy to use.

The evidence must be produced by the learner, without 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 5) 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 5)

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 are new to data science and those who wish to develop existing knowledge and skills (such as learners who have completed the corresponding level 4 unit). No previous knowledge of computer science, data science or statistics is assumed. However, learners should possess computational and numerical skills, which will be required to calculate statistics and perform the data analysis.

This unit is one in a series of units, with rising difficulty, that relate to data science. This is the second unit in the series. There is no requirement to undertake the units in sequence since each unit can be attempted without previous knowledge or experience of the subject.

The aim of the unit is to show learners what data science is, how it is used, and how to perform routine analyses on 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 large and varied, and include familiar and unfamiliar contexts. It is not appropriate to focus learning on small, familiar datasets. A critical objective 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.

Unit title: Data Science (SCQF level 5)

Although data ethics is not specifically included in the unit, the coverage of data bias (Outcome 1, Performance Criterion (c)) provides an opportunity to discuss this topic, which is covered more fully in the corresponding unit at level 6. It is recommended that the ethical implications of data science are emphasised throughout the unit.

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 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. However, the outcomes may be delivered holistically, without a clear delineation between them (see guidance on the content and context for this unit).

Outcome 1: This outcome explores the tools and techniques of data science. For those new to the subject, this will be their first exposure to the field so topics should be introduced with care. This outcome sets the scene for subsequent outcomes. Emphasis should be placed on the volume of data that is generated and the need for tools to harness this data.

The treatment of each topic should be relatively light. Breadth is more important than depth at this level. For example, in Performance Criterion (b), a wide range of applications of data science should be described in relatively shallow detail, rather than a narrow range of applications in depth. Emphasis should be placed on the application of data science in fields spanning astronomy to crime fighting.

The performance criteria are self-explanatory. Performance Criterion (d) provides an opportunity to introduce learners to a variety of tools to carry out analysis. For example, if learners use Excel™ to perform analyses, the use of alternative tools (such as Tableau™, Python and R) should be described (and combinations of these tools such as Excel libraries for Python). The limitations of generic tools (such as Excel™) should be explained.

Outcome 2: This outcome relates to methods of data analysis. Once again, given the level of this unit, treatment of any single topic should be light. For example, methods of cleaning and transforming data (Performance Criterion (c)) should be limited to the most common techniques for cleaning and structuring data. The key learning outcome is that learners appreciate the need for data cleaning and transforming, rather than an in-depth knowledge of the techniques for doing so.

The descriptive statistics (Performance Criterion (e)) should include a range of summary statistics including common measures of central tendency and dispersion such as sum, mean, median, mode, range, inter-quartile range and mean absolute deviation. Learners are required to describe these statistics, including worked examples.

Learners should be introduced to a range of visualisations (Performance Criterion (f)) and be able to select appropriate visualisations for different types of data. For example, learners should know the best types of visualisation to illustrate time series, statistical distributions, and comparisons and correlations between datasets.

Unit title: Data Science (SCQF level 5)

Outcome 3: This outcome applies the knowledge gained in Outcome 1 and Outcome 2. Learners are required to perform routines analyses on datasets. Learners should use relatively large datasets to explore the features of the software. Datasets with 1,000–5,000 records are recommended.

If this is 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 familiarity with the software and its analytical features. The terminology of data analytics (“clean”, “transform”, “visualise”, etc.) will require careful introduction.

At this level, learners are expected to understand and use specific analytical features of the software. For example, in the context of Excel™, learners would be expected to use the query, transform, pivot and visualisation features to gain insights in the data. The derived dashboards (Performance Criterion (f)) need not be particularly sophisticated but must be interactive.

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 methods in Outcome 2 and immediately practice those methods, using appropriate software, in Outcome 3. For example, once descriptive statistics are introduced in Outcome 2 (Performance Criterion (e)), learners can use software to calculate these statistics for a variety of small datasets (Outcome 3, Performance Criterion (d)).

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, in the context of population growth, learners could be supplied with a dataset showing demographics over time, and asked to address specific questions relating to this data. 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.

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 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 toolset.

Unit title: Data Science (SCQF level 5)

A suggested distribution of time is:

Outcome 1: 8 hours
Outcome 2: 12 hours
Outcome 3: 20 hours

If Outcome 2 and Outcome 3 are delivered holistically, then the combined time available to learn and practice data analysis would be 32 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 assignment for product evidence. The selected response test could comprise a multiple choice test for Outcome 1 and Outcome 2. The questions would relate to the identifications and descriptions defined in the performance criteria. The test would sample from the knowledge domain (Outcome 1 and Outcome 2). An appropriate pass mark would be set. The practical assignment would require learners to analyse a dataset and produce a data dashboard based on the dataset. A checklist could be used to assess the completed analysis.

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 identifications 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. 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.

Unit title: Data Science (SCQF level 5)

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

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, skills in using spreadsheet software are valued by employers.

This Unit has the Core Skill of Numeracy 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 Information and Communication Technology SCQF level 4 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 4 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 4 embedded. Core Skill of Numeracy SCQF level 5 embedded in this unit. The Critical Thinking component of Problem Solving at SCQF level 4 is embedded in this unit.	16/08/19

Additional copies of this unit specification can be purchased from the Scottish Qualifications Authority. Please contact the Business Development and Customer Support team, telephone 0303 333 0330.

General information for learners

Unit title: Data Science (SCQF level 5)

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 an introduction to data science. No previous knowledge or experience of data science or statistics is required.

Data science is becoming very important. Data science explores large amounts of data to identify patterns and trends and make predictions. For example, data science is used to help businesses make better decisions; data science is also used in sport to help teams analyse their performance.

It is expected that there will be many jobs in this area in the coming years. Everyone, 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 the tools and techniques of data science.
2 Introduction to data analysis.
3 Carrying out data analysis.

The introduction to the tools and techniques of data science explores how it is used in lots of areas such as astronomy, crime fighting and healthcare. This part of the unit also covers sources of public and private data. You will also understand how data science can go wrong because of bias.

The introduction to data analysis looks at 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 statistics that are commonly used in data science.

The final part of the unit looks at how you actually use computers to analyse data. You will use software (such as Microsoft Excel™) to carry out an analysis of a dataset. The datasets that you use will be large, perhaps up to 5,000 records. This part of the unit will give you practical skills in using data analysis software. You will learn how to create data dashboards to present information in an interactive and attractive manner.

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.

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 6. Alternatively, you could find out more about the impact of data science on society by undertaking *Data Citizenship* at SCQF level 5 or level 6, or dive deeper into specific aspects of data science such as machine learning or statistics by undertaking units in these areas.

General information for learners (cont)

Unit title: Data Science (SCQF level 5)

This Unit has the Core Skill of Numeracy 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 Information and Communication Technology SCQF level 4 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 4 is embedded in this unit. When a learner achieves the unit, their Core Skills profile will also be updated to include this component.