National Unit Specification

General information

Unit title: Data Science Project (SCQF level 6)

Unit code: J2GT 46

Superclass: CB

Publication date: June 2019

Source: Scottish Qualifications Authority

Version: 03, August 2019

Unit purpose

The purpose of this unit is to allow learners to complete an end-to-end data science project independently, using pre-existing knowledge and skills.

This capstone project will introduce the types of complex problems that can be solved using data science techniques, reinforce the learner's knowledge of data science and provide a meaningful opportunity to apply data science tools and techniques.

The unit is a **non-specialist** unit. Learner will require knowledge of the data science process, which could be evidenced by possession of J2GV 46 *Data Science* at SCQF level 6.

It is not expected that learners should have completed J2G2 45 *Data Science: Project* at SCQF level 5 prior to undertaking this unit. However, if they have undertaken the level 5 unit, much of the theory will already be familiar. The level of analysis, clarity of insights and depth of documentation expected for this unit is more stringent. Learners are required to carry out the project independently, to ensure their familiarity with all aspects of the process.

On completion of this unit, learners will have undertaken a complete data science project, from defining the problem in a project brief, analysing the collected data, evaluating and communicating the findings. The focus of the unit is on clarity and communication throughout each of the steps rather than the complexity of the analysis delivered.

Learners may progress to more specialised units in data science such as J2G6 46 *Machine Learning* at SCQF level 6.

National Unit Specification: General information (cont)

Unit title: Data Science Project (SCQF level 6)

Outcomes

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

1 Develop a project plan to address the problem.
2 Collect data safely and securely.
3 Analyse the collected data to extract insights.
4 Evaluate the analysis findings.

Credit points and level

1 National Unit credit point at SCQF level 6 (6 SCQF credit points at SCQF level 6)

Recommended entry to the unit

Whilst entry is at discretion of the centre, learners should possess basic knowledge of the data science process and skills in its application to small datasets, which could be evidenced by possession of J2G2 46 *Data Science* at SCQF level 6 and J2HN 46 *Data Citizenship* at SCQF level 6.

Core Skills

Achievement of this Unit gives automatic certification of the following:

Complete Core Skill Problem Solving 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, it should be delivered after the two mandatory units (*Data Science* and *Data Citizenship*) to permit learners to acquire the necessary knowledge and skills to attempt this unit.

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 Project (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

Develop a project plan to address the problem.

Performance criteria

(a) Define the problem being addressed
(b) Explain the background to the problem
(c) Explain potential benefits of solving the problem
(d) Compare different approaches to address the problem
(e) Select and justify a specific approach.
(f) Plan for the steps taken to address the problem

Outcome 2

Collect data safely and securely.

Performance criteria

(a) Identify the data required, and appropriate sources, for the chosen approach
(b) Assemble the required dataset.
(c) Employ secure methods for storing the data sample
(d) Recognise privacy considerations with the required data
(e) Identify sources of bias in the collected data sample
(f) Review the impacts of data quality issues

Outcome 3

Analyse the collected data to extract insights.

Performance criteria

(a) Prepare and pre-process data for analysis
(b) Calculate dataset summary metrics
(c) Interpret analysis to identify insights
(d) Create graphs to visualise insights

National Unit Specification: Statement of standards (cont)

Unit title: Data Science Project (SCQF level 6)

Outcome 4

Evaluate the analysis findings.

Performance criteria

(a) Interpret analysis to create insights
(b) Summarise insights to identify key findings
(c) Evaluate the findings
(d) Identify additional analyses that could be carried out in future

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 **one** form: product evidence.

The **product evidence** will relate to all outcomes. It will be produced individually, by each learner.

Learners will need to demonstrate that they understand the problem in the project brief and are able to choose an appropriate approach to address the problem. They will need to plan, collect and analyse the required data and then evaluate the effectiveness of the analysis. Each step of the process should be documented. Each learner should submit the following as the minimum evidence.

1 **Problem statement,** which defines the problem being addressed in the project brief and explain the context of the problem and benefits of addressing the problem.
2 **Project plan**, which specifies the milestones, timescales, data and storage approach.
3 **Summary of the dataset** required, collection method and storage approach.
4 **Analysis** of the dataset, including the steps taken and output of the analysis, with appropriate visualisations.
5 **Summary of key findings and future activities** from the analysis to be presented to a wider audience.

The scale and complexity of the problem should be consistent with the SCQF level of this unit (level 6) and involve the analysis of a dataset comprising at least 20,000 multi-variate records. The project must elicit evidence from learners that covers the key knowledge and skills in J2G2 46 *Data Science* at SCQF level 6.

National Unit Specification: Statement of standards (cont)

Unit title: Data Science Project (SCQF level 6)

The following level descriptors are particularly relevant to the evidence.

- Relate the subject to a range of practical applications.
- Apply knowledge and skills in using some of the routine practices and techniques in routine contexts that may have non-routine elements.
- Draw conclusions and suggest solutions.
- Produce and respond to detailed and relatively complex written and oral communication in both familiar and unfamiliar contexts.
- Use a wide range of numerical and graphical data in routine contexts, which may have non-routine elements.
- Take responsibility for carrying out a range of activities where the overall goal is clear, under non-directive supervision.

When possible, evidence should be a naturally occurring by-product of teaching and learning. This evidence must be produced **independently** by each learner. It may be produced over the life of the unit, under loosely controlled conditions (including access to reference materials). Authentication will be necessary. The guide to assessment provides further advice on methods of authentication.

The guidelines on approaches to assessment (see the support notes section of this specification) provides specific examples of instruments of assessment.

National Unit Support Notes

Unit title: Data Science Project (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 an optional unit within the NPA Data Science at SCQF level 6. It is intended to give learners the opportunity to use the skills, knowledge and understanding developed through the successful completion of the mandatory units within the award. It is recommended that this unit be delivered towards the end of the programme of study, when learners are in a position to demonstrate the knowledge and skills they have accumulated over the programme.

The general context for this unit is the rising use of data science, and its vocational relevance to an increasing number of occupations and professions. The unit is intended for non-specialists and may be offered as part of a wide range of programmes. No prior knowledge of computer science or statistics is required; however, numerical competency is presumed.

The unit aims to enable the learner to complete an end-to-end data science project from problem statement to delivering value independently. This value can come from either an implemented solution or insight that drives a different decision or approach. The focus of the unit is on completing the steps required for a successful data science project, with the outcomes aligned to these required steps. The learner is guided by these outcomes to demonstrate an understanding of the problem they are addressing and develop an approach to addressing it. Of lesser focus is the complexity of the analysis undertaken, although it is important to use analysis to develop solutions and communicate these.

Please note that the following guidance, relating to specific outcomes, does not seek to explain each performance criterion, which is left to the professionalism of the teacher. It 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 unit delivery. As such, it is not representative of the relative importance of each outcome or performance criterion.

Outcome 1: This outcome demonstrates that the learner fully understands the problem they are attempting to address. By investigating the wider context of the problem and identifying the benefits, this will allow the learner to align their analysis effort to the problem as a whole.

There is almost a limitless set of possibilities to what data can be used for, it could be to answer an interesting question, it could be to do something useful, or it could be to create something valuable. The best data projects aim to address all three areas.

National Unit Support Notes (cont)

Unit title: Data Science Project (SCQF level 6)

The kinds of questions that can be turned into data problems are:

- How much or how many?
- Which category
- Which group?
- Is this weird?
- Which option should be taken?

This outcome also demonstrates that the learner has considered a number of approaches for addressing the problem and can articulate the approach chosen from feasibility, complexity and timescale perspectives. The plan created can be high level but should demonstrate an understanding of the areas to be addressed including research, data capture, storage, analysis methods, tools, risks and stakeholders.

The types of questions to answer at this stage are:

- What data is needed?
- Where will this data come from?
- Is there access to the data?

Outcome 2: This outcome demonstrates that the learner has given consideration to the actual data and sample they are using. Data can be accessed/captured from a variety of sources: existing, internet/download/open data, manually captured through surveys, technology such as wearables and sensors. It is important that learners understand the importance of these steps prior to diving into the actual analysis.

The types of questions to answer at this stage are:

- Will there be enough data to provide a robust answer?
- Will there be too much data to store easily?
- Are there any social or moral implications of collecting or using this data?
- Are there any quality issues with the data being used?

Outcome 3: This outcome demonstrates the learner is able to carry out the actual analysis of the data. This can be done using whatever tools the learner is most comfortable with, from spreadsheets to open source or proprietary programming languages.

A key part of the analysis phase covers the pre-processing and manipulation of raw data into usable information. The types of activities expected to be undertaken cover:

- Importing — Reading the data into the chosen tool
- Tidying — Restructuring the data into a tidy dataset
- Minimising — Removing the variables that are not needed for the analysis
- De-duplicating/filtering/deletion — Removing unnecessary observations
- Formatting — Getting the data type correct
- Removing outliers/rescaling
- Merging — If data is in multiple datasets

The actual analysis is expected to cover activities such as metric creation, counts and frequency distributions, descriptive statistics, cross-tabulations and correlation scatterplots, but will be dependent on the problem statement.

Outcomes 4: This outcome demonstrates that the learner is able to make inferences from the analysis and communicate these to a wider audience, for example, a project stakeholder. This is a key skill to ensure that value is extracted from any data project. Consideration of the next steps is very important in the data science process as it provides a purpose to the project. Value can only be extracted from the project if a different decision is made, or an existing process is changed. This outcome ensures the learner reviews their insight and makes a conscious decision around what they will do with it next.

The types of questions to answer at this stage are:

- How does the data answer the original problem statement?
- How does the data help defend against any objections?
- Are there any limitations on the conclusions, or angles that have not been addressed?
- What are the conclusions?
- What could be done differently next time?
- What should happen next?

Guidance on approaches to delivery of this unit

The outcomes may be delivered in the order in which they are written. They have been written with a learning sequence in mind.

A suggested distribution of time, across the outcomes, is:

- Outcome 1: 8 hours
- Outcome 2: 6 hours
- Outcome 3: 16 hours
- Outcome 4: 10 hours

If learners are developing their own project briefs, then more time should be allowed for Outcome 1.

Support may be required to ensure safe capture and storage of data and access to any tools.

Although learners may be working on different project briefs, it is expected that a class will be working on each outcome at the same time.

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.

National Unit Support Notes (cont)

Unit title: Data Science Project (SCQF level 6)

Assessment evidence is required at all stages and outcomes and is expected to be in the form of an individual portfolio documenting the steps, decisions taken, plans, datasets, analysis and conclusions. It is recommended that for Outcome 4 a presentation is also prepared to support the communication of findings and next steps.

The project brief can be provided by either the assessor or by the learner, but the assessor must ensure that it meets the requirement of the unit. As this is an individual project, project briefs developed by the learner are encouraged. It is recommended that learners working on their own brief submit it to the assessor for approval before starting the project to make sure the brief is appropriate and relevant to the knowledge and skills to be assessed. If the same brief is given to a group of learners, then additional care should be taken with authentication. The learner may wish to use the evidence they have already produced in the group project for J2G4 46 *Data Science: Project* at SCQF level 5 in the individual project for this unit. It is encouraged, although it is recommended that the learner discusses this approach with the assessor before starting the project.

Formative assessment could be used to assess learners' knowledge at various stages throughout the life of the unit. An ideal time to gauge their knowledge would be at the end of each outcome. This assessment could be delivered through an item bank of selected response questions, providing diagnostic feedback to learners.

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

This Unit has the Core Skill of Problem Solving at 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.

In this unit the project may provide opportunities for learners to develop core skills in *Problem Solving*, *Communication*, *Numeracy*, *Information and Communication Technology (ICT)*.

Problem Solving (SCQF level 6):

- Critical Thinking is addressed in Outcomes 1 and 4
- Planning and organising are addressed in Outcome 1
- Review and evaluation are addressed in Outcome 4

Communication (SCQF level 6):

- Oral Communication is addressed in Outcome 4.
- Written Communication (writing) goes across the whole unit

National Unit Support Notes (cont)

Unit title: Data Science Project (SCQF level 6)

Numeracy (SCQF level 6):

- Using Graphical Information is addressed in Outcome 3
- Using Number is addressed in Outcomes 3 and 4

Information and Communication Technology (SCQF level 6):

- Accessing Information is addressed in Outcome 2
- Providing/Creating Information is addressed across the whole unit

The project will also provide opportunities for learners to develop skills in planning, decision-making, implementation, time management, evaluation, enterprise, employability and citizenship. The degree to which the opportunities are exploited will depend on the realism of the project brief chosen.

History of changes to unit

Version	Description of change	Date
02	Core Skill Problem Solving at SCQF level 5 embedded.	16/08/19
03	Unit codes changed due to hierarchy	23/08/19

General information for learners

Unit title: Data Science Project (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.

The purpose of this unit is to allow you to work independently to complete an end-to-end data science project. This will involve defining the problem in a project brief that you will address, understanding the importance of the problem, plan how you will carry out the project, collect the required data, deliver the analysis and communicate the findings. You will document the decisions and findings as you go, to create a portfolio of work that constitutes your data science project.

The unit is suitable if you have undertaken the mandatory units for the NPA Data Science (SCQF level 6) and will give you an opportunity to practice the skills and techniques you have learned on a real problem. It will provide the opportunity to further develop the Core Skills at SCQF level 6, such as *Problem Solving*, *Communications* and *Information and Communication Technology*, and other key skills in planning, decision-making, implementation, time management, evaluation, enterprise, employability and citizenship.

The problem you will be working on can be developed either by yourself or with support and guidance from your teacher. The data you will use can be existing, open source from the internet, manually captured through surveys, or created through technological tools such as sensors.

The key goal of this unit is to experience the steps involved in carrying out a data project, and to understand that planning and implementation are just as critical as the analysis phase. You will get an opportunity to evaluate your findings and present your next steps and implementation plan towards the end.

You may progress to more advanced *Data Science* units on completion of this unit if you wish to improve your knowledge and skills in this area.