Next Generation Higher National Unit Specification

Data Science (SCQF level 7)

Unit code: J68R 47

SCQF level: 7 (16 SCQF credit points)

Valid from: session 2022–23

Prototype unit specification for use in pilot delivery only (version 1.0) May 2022

This unit specification provides detailed information about the unit to ensure consistent and transparent assessment year on year.

This unit specification is for teachers and lecturers and contains all the mandatory information required to deliver and assess the unit.

This edition: May 2022 (version 1)

Unit purpose

The purpose of this unit is to introduce learners to data analysis, data visualisation and the communication of data. It provides learners with essential analysis skills and the ability to visualise and communicate data effectively.

This is a non-specialist unit, suitable for a wide range of learners. It is suitable for learners who wish to understand how to prepare and analyse data, and communicate the results of data analysis to facilitate data-driven decision-making. No previous experience is required, although numeracy skills and familiarity with computer software are assumed. Previous experience of spreadsheets is desirable but not essential.

Learners understand the data analysis process and how to use data analysis tools to extract, transform and analyse data. Basic statistical methods (such as correlation) are introduced in context. Learners gain insights into large datasets using a range of data analysis tools and techniques to facilitate data-driven decision-making.

Learners also develop practical skills in the design and creation of a range of data visualisations, and learn how to interpret visualisations created by others. They learn how to select visualisations for different datasets and use techniques for communicating data effectively and accurately. During the unit, learners use data visualisation software to create a variety of visual representations of data to provide insights and tell stories. You introduce learners to data ethics.

On completion of this unit, learners can carry out analyses of familiar data to aid decision-making. They can describe suitable forms of data communication for simple types of data, explain what makes effective data communication, prepare data for visualisation, interpret visualisations created by others, design and create data visualisations, and identify when information is presented in a misleading manner.

Learners may progress to the Data Science unit at SCQF Level 8.

Unit outcomes

Learners who complete this unit can:

1 extract data from different sources
2 transform data to prepare for analysis
3 analyse data to provide insights
4 interpret a range of data visualisations
5 explain the features of effective data communication
6 create a range of data visualisations for the results of data analysis

Evidence requirements

Learners must provide both knowledge and product evidence.

Knowledge evidence

Learners must produce knowledge evidence for all knowledge and/or skills statements for outcomes 4 and 5. They can produce evidence over an extended period under lightly controlled conditions. Learners need to produce the minimum amount of evidence to infer competence.

If you use a test, you can sample knowledge evidence. Learners must produce evidence under controlled conditions in terms of location, timing and access to reference materials. You must include both outcomes (at least partially) and most of outcomes 4 and 5 knowledge and skills statements in the sample, including:

- types of data visualisation
- data quality
- key elements of a data visualisation
- extraction of meaning from a data visualisation
- misleading representations

Knowledge evidence demonstrates that learners can:

- interpret at least three data visualisations that meet the following criteria:
  - — one of the visualisations must be a dashboard
  - — include visualisations for both numerical and categorical data
- identify the features of at least two data visualisations that are deliberately misleading and demonstrate that learners can:
  - — identify the misleading features
  - — describe at least one implication of the misleading visualisation

The data visualisations you use to gather knowledge evidence should allow learners to fully satisfy the evidence requirements stated above. You should either provide learners with at least five visualisations (three visualisations for interpretation and two misleading visualisations) or combine visualisations that amount to at least three visualisations. If you combine visualisations, they must contain enough data to allow learners to fulfil all the evidence requirements defined above.

Product evidence

Product evidence relates to outcomes 1, 2, 3 and 6 and consists of two parts.

Part 1: Produce at least one analysed dataset

The analysis must relate to real data and can be descriptive or diagnostic. The dataset should be extracted from at least two sources. It can be familiar to learners but must comprise at least 1000 multi-variate records, that require significant cleaning. The dataset must include a range of data types, including dates and text, some of which require categorisation prior to analysis.

Learners must carry out a range of transformations on the data including:

- resolving data types
- combining and splitting data
- removing unnecessary data.

The transformed dataset must be:

- properly structured
- suitable for analysis
- efficient in terms of data storage
- legal and ethical, with evidence that learners have considered data bias

The analysis of the dataset must include:

- at least one pivot table to summarise the data
- various measures of spread and dispersion, including standard deviation
- identification of all significant summaries and relationships in the data, including correlations where these exist
- insights into the dataset

Part 2: Produce at least three data visualisations, based on the analysed dataset(s) from Part 1

These visualisations should:

- include both numerical and categorical data
- demonstrate that an appropriate form of visualisation has been selected for the intended purpose, audience, and the data
- demonstrate appropriate graphical choices such as structure, aesthetics and annotations
- follow graphical conventions
- be accompanied by one or more statements that describe insights that learners intend the target audience to take away from the visualisation
- ensure that one must be a dashboard

Evidence can be produced over an extended period, under lightly controlled conditions and must be authenticated. The Guide to Assessment provides further advice on methods of authentication.

The SCQF level of this unit provides additional context on the nature of the required evidence and the associated standards. You should use the level descriptors (explicitly or implicitly) when making judgements about the evidence.

Knowledge and skills

The following table shows the knowledge and skills covered by the unit outcomes.

Knowledge	Skills
Learners should understand: ♦ internal and external data sources ♦ types of data (categorical and numerical data and their sub-types) ♦ common data formats ♦ data quality including data bias ♦ tools for extracting and transforming data ♦ types of data transformation ♦ common transformations including filtering, sorting, combining, separating and grouping ♦ data cleaning ♦ data loading ♦ types of software for data analysis ♦ descriptive and diagnostic analytics ♦ data structures including lists and tables ♦ data aggregation including pivot tables ♦ normal distribution ♦ summary statistics ♦ measures of central tendency ♦ measures of dispersion including standard deviation ♦ the concept of correlation and correlation co-efficient ♦ the significance of domain knowledge in data analysis ♦ the evolution of the visual representation of data ♦ the purpose of communicating with data ♦ forms of data communication including summaries, charts, infographics, dashboard and storytelling ♦ types of data visualisation ♦ key elements of a data visualisation including terminology and graphical conventions	Learners can: ♦ identify data sources ♦ extract data ♦ transform data ♦ clean and load data ♦ apply appropriate data analysis tools ♦ analyse data ♦ prepare data for visualisation ♦ identify key performance indicators ♦ design effective data visualisations ♦ match visualisation to audience ♦ create visualisations using software ♦ create dashboards and reports ♦ communicate insights to audiences

Knowledge	**Skills**
Learners should understand: ♦ extraction of meaning from a data visualisation ♦ design considerations ♦ selection of visualisations ♦ data dashboards and data storytelling ♦ how to facilitate understanding through appropriate representation and presentation ♦ misleading representations ♦ legal and ethical considerations	

Meta-skills

Throughout the unit, learners develop meta-skills to enhance their employability in the computing sector.

Self-management

This meta-skill includes:

- focusing: sorting, filtering
- integrity: ethics
- initiative: decision making

Social intelligence

This meta-skill includes:

- communicating: receiving information, giving information, storytelling
- leading: influencing

Innovation

This meta-skill includes:

- curiosity: information sourcing, problem recognition
- creativity: visualising
- sense-making: pattern recognition, synthesis, analysis
- critical thinking: deconstruction, logical thinking, computational thinking

Literacies

Throughout this unit, learners have opportunities to develop their literacy skills.

Numeracy

A wide range of numeracy skills are developed during this unit, particularly during the statistics topic.

Communication

Communication skills are developed when creating visualisations and communicating the results of analysis.

Digital

Digital skills are developed throughout this unit.

Delivery of unit

You should deliver outcomes in a sequential order. We suggest the following distribution of time:

Outcome 1 — Extract data from different sources
(10 hours)
Outcome 2 — Transform data to prepare for analysis
(15 hours)
Outcome 3 — Analyse data to provide insights
(15 hours)
Outcome 4 — Interpret a range of data visualisations
(10 hours)
Outcome 5 — Explain the features of effective data communication
(10 hours)
Outcome 6 — Create a range of data visualisations for the results of data analysis
(20 hours)

Learners require access to appropriate hardware and software for data analysis. You can use a range of software to facilitate learning (see 'Additional guidance' section). You should focus on the acquisition of practical skills in data analysis. However, you should introduce theory in context. For example, in outcome 3, the range of data formats should include a variety of common data formats, including those not likely to be used in learners' work roles. In the early stages of this unit, you have an opportunity to expose learners to the typical uses of data analysis using contemporary case studies.

When learners are in employment, you should give them the opportunity to work with datasets relevant to their work roles. Irrespective of their employment status, you should make use of familiar data that learners can relate to. The datasets used for learning should vary in their size and complexity. At some point in the unit, you should show learners realistic datasets in terms of size and messiness.

Both the interpretation and creation of data visualisations lend themselves to practical activities, so we suggest that as much of the delivery as possible is of a practical nature. For the interpretation of visualisations (outcomes 4 and 5), a large proportion of visualisations should be derived from real data. There are many visualisations openly available across many domains that you can use for this purpose, as well as open datasets to create visualisations for learners to interpret.

Activities and resources

Interpret a range of data visualisations (outcome 4)

- You can find examples of data visualisation in the media and current affairs in The Guardian, The New York Times, Chartr, Information is Beautiful and FiveThirtyEight.
- Column Five Media and Information is Beautiful both host good examples of infographics. 'The Functional Art' (Cairo, 2012) and 'The Truthful Art' (Cairo, 2016) are useful reference books.
- Geckoboard provides examples of corporate dashboards.

- ‘Fundamentals of Data Visualization’ (Wilke, 2019) is a useful reference e-book and ‘Data Visualisation: A Handbook for Data Driven Design’ (Kirk, 2016) is a comprehensive guide to data visualisation that contains an extremely useful guide to the major types of charts currently in use.
- For the use of storytelling in data communication, ‘Storytelling with Data’ (Knaflic, 2015) is a useful reference textbook. The data journalist Mona Chalabi mixes storytelling and infographics engagingly in the YouTube video Sequence, Sequence… Surprise! Designing Data for Maximum Impact.
- The blog post A Brief History of Data Visualization provides a useful historical summary of key milestones in the evolution of data visualisation.
- You can use games to show that it’s easier to interpret visual information than tables of numbers, for example speed games such as ‘spot the outlier’ or ‘spot the maximum’. Ask learners if it is easier to spot the feature of interest when it’s a number in a table or in a visualisation.
- When exploring the key elements of a chart, you can build up a chart one element at a time to help learners understand the importance of these elements. When does the chart become meaningful? When does it become useful? What are the elements of the chart that make it meaningful?
- When exploring graphical conventions, you can make links to conventions in literacy. How easy is it to understand a text when the author doesn’t follow spelling or grammatical conventions? You can make comparisons of reading a chart (data literacy) and reading a text.

Explain the features of effective data communication (outcome 5)

- You can find some examples of misleading charts on Statistics How To, and The Economist.
- Venngage provides a good summary of the common ways in which charts can be misleading (ironically, as an infographic).

Create a range of data visualisations for the results of data analysis (outcome 6)

- It is beneficial if learners peer-evaluate fellow learners’ visualisations and provide feedback. This gives a target audience for the learner creating the visualisation and reinforces the key elements of effective data communication.

Additional guidance

The guidance in this section is not mandatory.

Content and context for this unit

This unit introduces data analysis, data visualisation and communication to learners who have little or no experience in this field. It focuses on the practical aspects of analysing data, and the key knowledge and skills needed for communicating data. It also focuses on how to 'read' and make sense of the information contained in data visualisations and on the development of practical skills that enable learners to communicate the results of data analysis in ways that are accurate, trustworthy, accessible and useful.

You can use a variety of software tools for data analysis in this unit. We recommend that, if possible, you expose learners to a range of software. For example, you can deliver the entire unit using Microsoft Excel but it is best to do the data transformations using alternative software such as Query Editor. We recommend a combination of products.

Whenever possible, you should use real data for learning. This data should be familiar to learners and, ideally, relevant to their current work roles or vocational interests. Datasets should vary in their size and messiness but, at some stage in the unit, learners should experience large, messy datasets.

You should teach communicating with data after learners develop skills to extract, clean and transform data, and carry out analysis on it.

Extract data from different sources (outcome 1)

This outcome relates to data extraction, and the knowledge and skills statements are self-explanatory.

Data quality is an important part of this outcome. It is vital that learners appreciate the importance of data quality in data analysis. In discussion about data bias ('bias in' and 'bias out'), you should emphasise the problem of historical data bias and include ways of addressing this problem.

Transform data to prepare for analysis (outcome 2)

This outcome relates to data transformation. You should emphasise the importance and time-consuming nature of this stage in the data analysis process. At this level, the required transformations should be non-complex, but learners should gain experience of cleaning large, messy datasets.

This outcome also covers the legal and ethical aspects of data analysis. With regards to ethics, you should make learners aware of the difficulty of protecting the privacy of data subjects due to the inherent problem of ensuring anonymity in datasets. You should explore methods of improving data privacy (such as differential privacy). There is scope in this outcome to discuss the tension between privacy and accuracy — how increasing privacy (through anonymising the data or not recording certain attributes such as race or gender) can reduce the descriptive or predictive accuracy of the analysis.

Analyse data to provide insights (outcome 3)

This outcome relates to data analysis. Statistical concepts and methods should be basic. The most complex techniques are standard deviation and correlation, however, you can limit correlation to the Pearson correlation co-efficient(s). It is important that learners understand the underlying statistical concepts and not merely compute statistical values. They should understand the distinction between descriptive analytics and diagnostic analytics. The statistical topics covered can include:

- mean, median and mode
- range and interquartile range
- outliers
- percentages and percentiles
- normal distribution
- standard deviation
- types of growth, including linear and exponential
- correlation, including confounding variables
- misleading statistics

Interpret a range of data visualisations (outcome 4)

You should provide learners with a brief introduction to how data visualisation has evolved over time from the early inventors (Playfair, Nightingale, Minard and Snow) through more recent innovators (Tufte and Wilkinson) to modern-day examples. The purpose of this introduction is to show that people have found data visualisation valuable for hundreds of years, and that, while new types of visualisations are being created all the time, the 'core' charts have stood the test of time.

You could introduce learners to the interpretation of data visualisations through their use in the media and in current affairs. Data visualisation is increasingly common-place and the need to be able to interpret charts and other forms of data visualisation is becoming increasingly important for us as individuals, consumers and informed citizens.

You should introduce learners to how data communication is used to enable individuals and organisations to make informed decisions in a variety of domains, such as business, science and sport. It should be possible to find examples that match the interests and/or experience of learners. For a variety of domains, you should provide examples of the kinds of questions that can be answered by using data visualisations. Additionally, you should provide examples of how data visualisations can prompt questions which invite further investigation and better decision-making.

You should make the place of data communication in a data science process clear, with a particular focus on the relationship between data analysis and data communication.

When demonstrating to learners how the human brain can spot patterns, trends and outliers more easily in a data visualisation than in a table containing raw data you should do this in a practical manner.

You should expose learners to good examples from each of the following forms of data communication:

- **Summaries:** Descriptive statistics that summarise data in a single number, for example measures of central tendency for continuous data or proportions and percentages for discrete data. You should explore the advantages and limitations of summarising data using a single number.
- **Infographics**: You should make the distinction between infographics and charts; infographics may include charts but also typically include other graphical design elements, such as illustrations, photographs and diagrams, as well as textual information. They tend to be design-focused, static and are commonly used in a marketing context.
- **Dashboards**: At the time of writing, dashboards are a popular means to communicate data in a business context, particularly for data that relates to key performance indicators (KPIs). They typically comprise multiple summary statistics and charts, and are often interactive, enabling the data to be explored. You should show learners examples of dashboards used within different business functions.
- **Charts**: The term ‘chart’ can also mean ‘graph’ or ‘plot’, for example graphic representations of data. The charts that you should include in this unit are bar charts (and their sub-types — horizontal or vertical, stacked and grouped), histograms, scatter plots and line charts. Charts that show two-dimensional data on two axes are adequate and you do not have to demonstrate charts that show more than two axes. A brief examination of visualising dimensions beyond two dimensions (such as using colour, shape, size, facets or animation) is desirable. You can find an example of a scatter plot that uses size and colour to visualise four dimensions on the Gapminder website. You can include pie charts as examples of commonly used charts that have a better alternative.
- **Storytelling**: Refers to the construction of a verbal or written narrative structure that complements one or more data visualisations to make the ‘message’ conveyed by the visualisation(s) more compelling.

The recommended distribution of time that you spend on each form of data communication within this unit is as follows:

- Summaries: 5 per cent
- Infographics: 5 per cent
- Dashboards: 10 per cent
- Charts: 70 per cent
- Storytelling: 10 per cent

For each chart type, learners should know the names and purposes of the key elements of the chart, such as title, axes, labels, and legend. They should be able to:

- use the correct terminology
- identify the elements required to make the chart meaningful
- identify data that is ready to be communicated

Learners should appreciate the role of graphical conventions (such as numbers on axes starting at zero, or spacing tick marks at regular intervals) in making charts easier to consume.

For outcome 5, 'Explain the features of effective data communication', learners should be aware of the various aspects of design that make effective data visualisation easier, namely the choice of data, structure (such as co-ordinate system and scales), annotations and aesthetics (such as colour and shape).

Learners should explore how bad design choices slow understanding and good design choices help it. They should be able to select an appropriate form of data communication based on the:

- goal of the communicator
- purpose of the communication
- intended audience
- method of consumption (for example, will it be delivered in a face-to-face presentation or embedded within a written document?)

Learners can select from summary, infographic, dashboard, or chart forms of data communication. You should encourage them to select charts or dashboards for this unit, in that order.

When learners select a chart as the appropriate form of communication, they should be able to select the appropriate representation (such as chart type) for the type of data being visualised.

In terms of audience awareness, learners should consider what they want to get as recipients of the communication, what their experience of data visualisation is, and what their needs are (for example accessibility needs such as colour blindness). You should show them a range of examples of misleading charts, including some that are purposefully misleading and some that are simply badly designed. You can include problematic pie charts, such as those with many segments, or angled 3D pie charts where the area of each segment is difficult to determine. You should give learners the opportunity to identify the elements of the charts that cause them to be misleading, and what the creator of the chart would need to do to address the issue(s).

You should teach learners about data bias and give them the opportunity to explore the reasons why some people deliberately create misleading charts (to influence others) and the ethical implications of this. You can draw examples from politics or marketing, amongst other areas. Learners should also explore ways in which bias can be reduced. This should give them the opportunity to explore the importance of producing data visualisations that are accurate and trustworthy.

Create a range of data visualisations for the results of data analysis (outcome 6)

Using the skills developed in outcomes 4 and 5, learners should be able to select an appropriate visualisation to create, based on the goals and purpose of the communication, the audience and the data, and make appropriate design choices about this visualisation. It is beneficial if a variety of contexts and purposes are used to create the visualisations.

The choice of software used to create the visualisations is at learners and your discretion, but should not require programming.

Using a software package, learners should create the following charts from a variety of simple datasets: bar charts (and their sub-types — horizontal or vertical, stacked and grouped), histograms, scatter plots and line charts. Using a software package, learners should create at least one dashboard.

Using what they have learned about effective data communication in outcome 5, learners should be able to evaluate what they have produced, and take remedial action to address any issues, before communicating the insights obtained from their visualisations.

Approaches to assessment

Evidence can be generated using different types of assessment. The following are suggestions only, as there may be other methods that would be more suitable to learners.

Outcomes 4 and 5

We suggest gathering the knowledge evidence required for outcomes 4 and 5 with a written test consisting of a mix of multiple-choice, restricted response or short answer questions, marked and assessed traditionally, where the questions are based on:

- at least three data visualisations that meet the following criteria:
  - one of the visualisations is a dashboard
  - includes visualisations for both numerical and categorical data
  - can be a simple bar chart or scatter plot
  - can be accompanied by an explanation describing the purpose and context in which the data visualisation was produced
  - can be accompanied by a brief, the purpose of which is to set one or more specific goals or define one or more specific questions which the visualisation should be used to answer
  - can be derived from real or fictitious data
- at least two data visualisations that are deliberately misleading

The test must take place under supervised and closed-book conditions. Learners cannot bring any notes, textbooks, handouts or calculators to the test and they should complete the test within 1 hour and 30 minutes. The questions presented must change on each test occasion. Learners should answer at least 60 per cent of the questions correctly to pass this test.

If you are presenting this test online, you can select the following question types, where appropriate:

- multiple-choice
- drag-and-drop
- multiple response
- mix-and-match
- a combination of the above

Alternatively, for outcome 5, you can gather the knowledge evidence through oral questions based on the use of at least two deliberately misleading data visualisations as stimuli. As a form of formative assessment this could take place during a shared discussion on the ethical dimensions of misleading data visualisation and its implications.

A more contemporary approach to assessment involves the use of a blog to record learning and researched case study examples throughout the life of the unit. The blog would provide knowledge evidence in the descriptions and explanations. You should assess the blog using defined criteria, to permit a correct judgement about the quality of the evidence. In this scenario, every knowledge and skill must be evidenced; sampling is not appropriate.

Outcomes 1, 2, 3 and 6

The evidence requirements for the practical outcomes can be satisfied in two ways. Towards the end of the unit, you can set an assignment that requires learners to carry out a specific analysis on a familiar dataset and produce data visualisations for communication. You can use a case study. This practical assignment must satisfy all the evidence requirements. Alternatively, learners can maintain a portfolio of activities carried out during the life of the unit. The portfolio would collectively satisfy all the evidence requirements, but at least one analysis must comprise at least 1000 records.

If using a case study, you should provide a clear brief. This should relate to a specific data analysis, visualisation and communication tasks faced by a team within a particular organisation, have a specific objective, and be targeted at a particular audience. It should include background information on the organisation, the task, the team and the audience.

You can develop a checklist that defines the standards to be achieved.

Learners can choose the method of communicating the data visualisations they produce. For example, this could be by embedding them in a written report, presenting them orally, or publishing them as a blog.

Equality and inclusion

This unit is designed to be as fair and as accessible as possible with no unnecessary barriers to learning or assessment.

You should take into account the needs of individual learners when planning learning experiences, selecting assessment methods or considering alternative evidence.

Guidance on assessment arrangements for disabled learners and/or those with additional support needs is available on the assessment arrangements web page: www.sqa.org.uk/assessmentarrangements.

Information for learners

Data Science (SCQF Level 7)

This section explains:

- what the unit is about
- what you should know or be able to do before you start
- what you need to do during the unit
- opportunities for further learning and employment

Unit information

This unit introduces you to data analysis, visualisation and communication of data for non-specialists. You learn from the beginning. No previous knowledge of data analysis, data visualisation or computer programming is assumed. A familiarity with spreadsheet software is desirable but not essential.

You use a variety of software to carry out your analyses, such as Microsoft Excel and Query Editor. You use this software to practise with real data to answer real business questions. This unit also helps you make sense of the range of information that we are all exposed to in a graphical format and introduces you to the interpretation of information presented in this way. Examples are graphs, infographics, charts and plots, and these can all be termed data visualisation. You may have seen examples of graphs where the data is presented in a misleading way. This unit helps you to identify those creating graphs are trying to mislead you and explores the ethics around producing accurate or misleading visualisations.

During the unit, you work with datasets relevant to your current work role. You learn how to use data analysis to gain insights into these datasets and make data-driven decisions based on your analysis. You also learn:

- how individuals and organisations use data visualisations to make informed decisions
- how to interpret a range of data visualisations including bar charts, histograms, scatter plots and line charts
- how to identify misleading information in visualisations
- what you need to do to communicate data effectively
- how to create your own data visualisations to provide insights

This unit covers a variety of topics including:

- the ETL (extract, transform and load) process
- types of data
- data formats
- data quality
- data transformations
- data cleaning

- data structures
- basic statistics
- descriptive and diagnostic analytics
- using software to perform these analyses
- purposes of data communication
- forms of data communication
- types of data visualisation
- how data visualisation has evolved
- extracting meaning from a data visualisation
- key considerations for good data communication
- how good representation and presentation choices facilitate understanding
- selecting the right type of visualisation to use
- designing data visualisations
- preparing data for visualisation
- creating visualisations using software
- ways in which data visualisation can mislead
- data ethics

You may be assessed using a range of methods, most of which are practical in nature, including practical assignments or case studies.

This unit also provides opportunities for you to enhance your meta-skills in self-management, social intelligence and innovation.

After completing this unit, you may progress to:

- Data Science at SCQF level 8.

Administrative information

Published: May 2022 (version 1.0)

Superclass: CB

History of changes

Version	Description of change	Date

Note: please check SQA's website to ensure you are using the most up-to-date version of this document.