Data Scientist

Certification Specification (Version 2024.1)

Certification Purpose

This certification verifies that individuals have acquired the knowledge and skills required of entry-level data scientists.

The awarded certification will be valid for two years.

Target Audience

The Data Scientist Certification is intended to be used by students and potential employers for demonstrating the level of skill required for employment into an entry-level role. Specific stakeholders are:

- **Data Scientists** aiming to demonstrate that they have acquired the skills required for an entry-level position through a course of training/education
- **Hiring managers** wanting to confirm the skill level of job applicants
- **Students** wanting to measure their achievement in a learning programme
- **Employers** wanting to confirm the current skill level of their teams

Domains and Competencies

In addition to the competencies required for the associate level (see Data Science Associate Specification), Data Scientists will be required to demonstrate competency in five domains:

- Data Management
- Exploratory Analysis
- Programming for Data Science
- Data Communication
- Business Acumen

The following sections break down each domain into the detailed knowledge, skills and abilities (KSAs) that will be required to be demonstrated, along with technologies required where relevant.

Data Management

1. Collect data from non-standard formats (e.g. json) by modifying existing code (Python or R)

    1.1. Adapt provided code to import data from an API using Python or R.

    1.2. Identify the structure of HTML and JSON data and parse them into a usable format for data processing and analysis using Python or R.

2. Perform data extraction, joining and aggregation tasks (SQL)

    2.1. Aggregate numeric, categorical variables and dates by groups using PostgreSQL.

    2.2. Interpret a database schema and combine multiple tables by rows or columns using PostgreSQL.

    2.3. Extract data based on different conditions using PostgreSQL.

    2.4. Use subqueries to reference a second table (e.g. a different table, an aggregated table) within a query in PostgreSQL

Exploratory Analysis

1. Identify and reduce the impact of characteristics of data

    1.1. Identify when imputation methods should be used and implement them to reduce the impact of missing data on analysis or modeling using R or Python.

    1.2. Describe when a transformation to a variable is required and implement corresponding transformations using R or Python.

    1.3. Describe the differences between types of missingness and identify relevant approaches to handling types of missingness.

    1.4. Identify and handle outliers using R or Python.

Programming for Data Science

1. Demonstrates best practices in production code including version control, testing and package development (Python)

    1.1. Describe the basic flow and structures of package development

    1.2. Explain how to document code in packages, or modules

    1.3. Explain the importance of the testing and write testing statements

    1.4. Explain the importance of version control and describe key concepts of versioning

Data Communication

1. Frame, convey, and summarize stories using data

    1.1. Employ techniques in data storytelling to propose findings and relay solutions to business stakeholders

2. Employ multiple tactics (written and verbal) to communicate to business leaders

    2.1. Deliver a verbal presentation addressing the business goals, outcomes and recommendations

    2.2. Provide a written explanation of findings and/or reasoning for selecting approaches

Business Acumen

1. Make recommendations for analytic approaches based on business goal

    1.1. Explain how solution addresses the business problem

    1.2. Provide recommendations for future action to be taken based on the outcome of the work done

2. Judge performance of analytic results against relevant business criteria

    2.1. Define a KPI to compare model performance to business criteria in the problem

    2.2. Compare the performance of the two models/approaches using the defined KPI

Exam Format

To earn the Data Scientist Certification, candidates will have:

- To complete 3 exams, or 2 exams if they currently hold the associate level certification
- A total of 30 days to complete all of the exams
- To pass the exams in the order specified
- To pass each exam before progressing to the next
- Two attempts to pass each exam within the 30 days

All exams are offered in English only.

Exams will be taken online. Candidates will need a laptop or computer with internet access.

Exam	Type	Number of items	Time Available

DS101	Timed	45	2 hours
DS201	Timed	60	2 hours
DS601P	Practical	NA	NA

Or for holders of the Data Science Associate Certification:

Exam	Type	Number of items	Time Available
DS202	Timed	60	2 hours
DS601P	Practical	NA	NA

Results

Results for the timed exams will be displayed upon completion of the exam. Results will include a score for each domain as well as an overall score. The overall score determines the final pass/fail decision.

Results for the timed exams are calculated as an average of the scores for each domain included in the exam.

Results for the practical exam will be available within 14 days of submission. Candidates will be emailed to inform them that their results are available to review. The practical exam will be a pass/fail exam. All criteria must be passed. No scores will be given.

Re-taking

Candidates who are unsuccessful in any component will have to wait 14 days before they can attempt the certification again. They will be informed of the domains where they were unsuccessful. They will have to complete all exams again, including any that they may have passed on a previous attempt.

Unscored Items

Candidates may be presented with one unscored item for each domain & technology combination during a timed exam. These items are included to validate them for use in future exams. They will not count towards the final score received. They will not be highlighted during the test taking process.

Exam Content

The following sections outline the content breakdown for each exam. Each section contains an overview of the number of items that we will ask for each domain within an exam. There is also a breakdown of the overall pool size and the split of the types of questions contained within each domain.

DS101

The first exam for the Data Science Certification will cover some of the competencies from the associate level (please refer to the associate level specification for the detailed competencies and KSAs).

This exam contains the Statistical Experimentation Python or R and theory content as well as the exploratory analysis Python or R content.

Candidates will have to select either Python or R before starting the exam and both domains will use the same technology.

Domain	Technology	Number of competencies	Items per Competency	Total Items in Exam
Statistical Experimentation	Python or R	2	7-8	15
Statistical Experimentation	Theory	1	15	15
Exploratory Analysis	Python or R	4	3-5	15
				45

DS201

The second exam for the Data Scientist Certification will cover all of the remaining domains and competencies.

Candidates will have to select either Python or R before starting the exam and all domains will use the same technology.

Domain	Technology	Number of competencies	Items per Competency	Total Items in Exam
Data Management	Python or R	4	3-5	15
Data Management	SQL	1	15	15
Programming for Data Science	Python or R	2	7-8	15
Model Development	Python or R	4	3-5	15

				60

DS202

This exam is intended for anyone who already holds the associate level Data Science certification. It will cover all of the domains and competencies not yet tested at associate level.

Candidates will have to select either Python or R before starting the exam and all domains will use the same technology.

Domain	Technology	Number of competencies	Items per Competency	Total Items in Exam
Data Management	Python or R	4	3-5	15
Data Management	SQL	1	15	15
Programming for Data Science	Python or R	1	15	15
Exploratory Analysis	Python or R	4	3-5	15
				60

DS601P

The practical exam will test KSAs in the following domains:

- Data Management
- Exploratory Analysis
- Model Development
- Data Communication
- Business Acumen

The exam must be completed in Python or R.

Candidates will be given a single real-world scenario. They will be required to:

- submit a written report intended for a technical audience
- deliver a spoken presentation intended for a non-technical audience

Submissions will be graded by human graders against the following criteria:

Competency	Sufficient	Insufficient

Data Management		
Assess data quality and perform validation tasks	Has validated all variables and where necessary has performed cleaning tasks to result in analysis-ready data.	Has not conducted all the required checks and/or has not cleaned the data. May have removed data rather than performed cleaning tasks.
Exploratory Analysis		
Create data visualizations in R or Python to demonstrate the characteristics of data and represent the relationships between features	Has created at least two different visualizations of single variables (e.g. histogram, bar chart, single boxplot) Has created at least one visualization including two or more variables (e.g. scatterplot, filled barchart, multiple boxplots) Has used visualizations that support the findings being presented	Has used the same visualization throughout. Has not included graphics to represent single variables and relationships. Has not used visualizations that support the findings being presented.
Model Development		
Implement standard modeling approaches for supervised learning problems	Correctly identified the type of problem (regression, classification or clustering) Has selected and fitted a model for that problem to be used as a baseline. Has selected and fitted a comparison model for the problem that they were provided.	Has incorrectly identified the type of problem. Has not fitted a baseline model or has used a model for the wrong type of problem. Has not fitted a comparison model or has used a model for the wrong type of problem.
Use suitable methods to assess the performance of a model	Compared the performance of the two models/approaches using any method appropriate to the type of problem. Has described what the model comparison shows about the selected approaches.	Has selected a method not suitable for the type of problem. Has not described what the results show about the selected approaches.

Data Communication		
Employ multiple tactics (written and verbal) to communicate to business leaders	For each analysis step, has provided a written explanation of their findings and/or reasoning for selecting approaches Has delivered a verbal presentation addressing the business goals, outcomes and recommendations	Has not provided a written summary for each step Has not delivered a verbal presentation
Business Acumen		
Make recommendations for analytic approaches based on business goal	Has described at least one of the business goals of the project Has explained how their work has addressed the business problem Has provided at least one recommendation for future action to be taken based on the outcome of the work done	Has not identified any business goals Has not explained how their work has addressed the business problem Has not provided any recommendations for future actions
Judge performance of analytic results against relevant business criteria	Has defined a KPI to compare model performance to business criteria in the problem Has compared the performance of the two models/approaches using the defined KPI	Has not identified a KPI to compare the model performance to the business problem Has not compared the performance of the two approaches using the defined KPI

Item Formats

The following information provides examples of the item formats used across the timed and practical exams and their purpose.

Timed Exams

The timed exams will use items of three formats:

- Multiple choice

- Fill-in-the-blank
- Typing

Practical Exam

The practical exam will require submission of a published workbook and recording of a presentation.

Test takers will need to create a workbook from their Certification dashboard that will be shared with members of the grading team and DataCamp Certification team admins for grading purposes.

On submission of their published workbook, candidates will be able to use our recording tool to record their submission. It is not required to submit presentation materials. It is not possible to submit recordings in any other format.

Submissions will be automatically allocated for grading at random to a member of the grading team. A proportion of submissions will be allocated to multiple graders as part of our quality assurance process.

An example of a practical exam can be viewed here.