Data Scientist Certification Study Guide

Please use this study guide to create your certification self-study plan. We've included the objectives you should meet for each assessed competency, with links to relevant practice assessments.

- **Associate Certification**
  - Exams DS101 and DS102
- **Professional Certification**
  - Exams DS101 and DS201

Associate and Professional

Exam DS101: Exploratory Analysis and Statistical Experimentation in R or Python

1.1 Calculate metrics to effectively report characteristics of data and relationships between features

- Calculate measures of center (e.g. mean, median, mode) for variables using R or Python.
- Calculate measures of spread (e.g. range, standard deviation, variance) for variables using R or Python.
- Calculate skewness for variables using R or Python.
- Calculate missingness for variables and explain its influence on reporting characteristics of data and relationships in R or Python.
- Calculate the correlation between variables using R or Python.

1.2 Create data visualizations in coding language to demonstrate the characteristics of data

- Create and customize bar charts using R or Python.
- Create and customize box plots using R or Python.
- Create and customize line graphs using R or Python.
- Create and customize histograms graph using R or Python.

1.3 Create data visualizations in coding language to represent the relationships between features

- Create and customize scatterplots using R or Python.
- Create and customize heatmaps using R or Python.
- Create and customize pivot tables using R or Python.

Data Scientist Certification Study Guide

1.4 Identify and reduce the impact of characteristics of data

- 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.
- Describe when a transformation to a variable is required and implement corresponding transformations using R or Python.
- Describe the differences between types of missingness and identify relevant approaches to handling types of missingness.
- Identify and handle outliers using R or Python.

Related Assessments

Data Manipulation with R

Data Manipulation with Python

2.1 Describe statistical concepts that underpin hypothesis testing and experimentation

- Define different statistical distributions (e.g. binomial, normal, Poisson, t-distribution, chi-square, and F-distribution, etc.).
- Explain the statistical concepts in hypothesis testing (e.g. null hypothesis, alternative hypothesis, one-tailed and two-tailed hypothesis tests, etc.).
- Explain the statistical concepts in the experimental design (e.g. control group, randomization, confounding variables, etc.).
- Explain parameter estimation and confidence intervals.

2.2 Apply sampling methods to data

- Distinguish between different types of random sampling techniques and apply the methods using R or Python
- Sample data from a statistical distribution (e.g. normal, binomial, Poisson, exponential, etc.) using R or Python
- Calculate a probability from a statistical distribution (e.g. normal, binomial, Poisson, exponential, etc.) using R or Python

2.3 Implement methods for performing statistical tests

- Run statistical tests (e.g. t-test, ANOVA test, chi-square test) using R or Python.
- Analyze the results of statistical tests from R or Python.

Data Scientist Certification Study Guide

Related Assessments

Statistics Fundamentals with R

Statistics Fundamentals with Python

Associate Only

Exam DS102: Data Management, Modeling, and Programming in R or Python

2.1 Perform standard data import, joining and aggregation tasks

- Import data from flat files into R or Python.
- Import data from databases into R or Python
- Aggregate numeric, categorical variables and dates by groups using R or Python.
- Combine multiple tables by rows or columns using R or Python.
- Filter data based on different criteria using R or Python.

2.2 Perform standard cleaning tasks to prepare data for analysis

- Match strings in a dataset with specific patterns using R or Python.
- Convert values between data types in R or Python.
- Clean categorical and text data by manipulating strings in R or Python.
- Clean date and time data in R or Python.

2.3 Assess data quality and perform validation tasks

- Identify and replace missing values using R or Python.
- Perform different types of data validation tasks (e.g. consistency, constraints, range validation, uniqueness) using R or Python.
- Identify and validate data types in a data set using R or Python.

2.4 Collect data from non-standard formats by modifying existing code

- Adapt provided code to import data from an API using R or Python.
- Identify the structure of HTML and JSON data and parse them into a usable format for data processing and analysis using R or Python.

Related Assessments

Importing and Cleaning with R

Importing and Cleaning with Python

Data Scientist Certification Study Guide

3.1 Prepare data for modeling by implementing relevant transformations.

- Create new features from existing data (e.g. categories from continuous data, combining variables with external data) using R or Python.
- Explain the importance of splitting data and split data for training, testing, and validation using R or Python.
- Explain the importance of scaling data and implement scaling methods using R or Python.
- Transform categorical data for modeling using R or Python.

3.2 Implement standard modeling approaches for supervised learning problems.

- Identify regression problems and implement models using R or Python.
- Identify classification problems and implement models using R or Python.

3.3 Implement approaches for unsupervised learning problems.

- Identify clustering problems and implement approaches for them using R or Python.
- Explain dimensionality reduction techniques and implement the techniques using R or Python.

3.4 Use suitable methods to assess the performance of a model.

- Select metrics to evaluate regression models and calculate the metrics using R or Python.
- Select metrics to evaluate classification models and calculate the metrics using R or Python.
- Select metrics and visualizations to evaluate clustering models and implement them using R or Python.

Related Assessments

Machine Learning Fundamentals in R

Machine Learning Fundamentals in Python

4.1 Use common programming constructs to write repeatable production quality code for analysis.

- Define, write and execute functions in R or Python.
- Use and write control flow statements in R or Python.
- Use and write loops and iterations in R or Python.

Data Scientist Certification Study Guide

4.2 Demonstrates best practices in production code including version control, testing, and package development.

- Describe the basic flow and structures of package development in R or Python.
- Explain how to document code in packages, or modules in R or Python.
- Explain the importance of the testing and write testing statements in R or Python.
- Explain the importance of version control and describe key concepts of versioning

Related Assessments

R Programming

Python Programming

Professional Only

Exam DS201: Data Management in SQL; Data Management, Modeling, and Programming in R or Python

1.1 Perform data extraction, joining and aggregation tasks

- Aggregate numeric, categorical variables and dates by groups using PostgreSQL.
- Interpret a database schema and combine multiple tables by rows or columns using PostgreSQL.
- Extract data based on different conditions using PostgreSQL.
- Use subqueries to reference a second table (e.g. a different table, an aggregated table) within a query in PostgreSQL

Related Assessment

Data Management in SQL (PostgreSQL)

2.1 Perform standard data import, joining and aggregation tasks

- Import data from flat files into R or Python.
- Import data from databases into R or Python
- Aggregate numeric, categorical variables and dates by groups using R or Python.
- Combine multiple tables by rows or columns using R or Python.
- Filter data based on different criteria using R or Python.

2.2 Perform standard cleaning tasks to prepare data for analysis

Data Scientist Certification Study Guide

- Match strings in a dataset with specific patterns using R or Python.
- Convert values between data types in R or Python.
- Clean categorical and text data by manipulating strings in R or Python.
- Clean date and time data in R or Python.

2.3 Assess data quality and perform validation tasks

- Identify and replace missing values using R or Python.
- Perform different types of data validation tasks (e.g. consistency, constraints, range validation, uniqueness) using R or Python.
- Identify and validate data types in a data set using R or Python.

2.4 Collect data from non-standard formats by modifying existing code

- Adapt provided code to import data from an API using R or Python.
- Identify the structure of HTML and JSON data and parse them into a usable format for data processing and analysis using R or Python.

Related Assessments

Importing and Cleaning with R

Importing and Cleaning with Python

3.1 Prepare data for modeling by implementing relevant transformations.

- Create new features from existing data (e.g. categories from continuous data, combining variables with external data) using R or Python.
- Explain the importance of splitting data and split data for training, testing, and validation using R or Python.
- Explain the importance of scaling data and implement scaling methods using R or Python.
- Transform categorical data for modeling using R or Python.

3.2 Implement standard modeling approaches for supervised learning problems.

- Identify regression problems and implement models using R or Python.
- Identify classification problems and implement models using R or Python.

3.3 Implement approaches for unsupervised learning problems.

- Identify clustering problems and implement approaches for them using R or Python.

Data Scientist Certification Study Guide

- Explain dimensionality reduction techniques and implement the techniques using R or Python.

3.4 Use suitable methods to assess the performance of a model.

- Select metrics to evaluate regression models and calculate the metrics using R or Python.
- Select metrics to evaluate classification models and calculate the metrics using R or Python.
- Select metrics and visualizations to evaluate clustering models and implement them using R or Python.

Related Assessments

Machine Learning Fundamentals in R

Machine Learning Fundamentals in Python

4.1 Use common programming constructs to write repeatable production quality code for analysis.

- Define, write and execute functions in R or Python.
- Use and write control flow statements in R or Python.
- Use and write loops and iterations in R or Python.

4.2 Demonstrates best practices in production code including version control, testing, and package development.

- Describe the basic flow and structures of package development in R or Python.
- Explain how to document code in packages, or modules in R or Python.
- Explain the importance of the testing and write testing statements in R or Python.
- Explain the importance of version control and describe key concepts of versioning

Related Assessments

R Programming

Python Programming