Module 1 Summary

Congratulations! You have completed this module. At this point in the course, you know:

- The Data Science Task Categories include:
    - Data Management - storage, management and retrieval of data
    - Data Integration and Transformation - streamline data pipelines and automate data processing tasks
    - Data Visualization - provide graphical representation of data and assist with communicating insights
    - Modelling - enable Building, Deployment, Monitoring and Assessment of Data and Machine Learning models
- Data Science Tasks support the following:
    - Code Asset Management - store & manage code, track changes and allow collaborative development
    - Data Asset Management - organize and manage data, provide access control, and backup assets
    - Development Environments - develop, test and deploy code
    - Execution Environments - provide computational resources and run the code

The data science ecosystem consists of many open source and commercial options, and include both traditional desktop applications and server-based tools, as well as cloud-based services that can be accessed using web-browsers and mobile interfaces.

Data Management Tools: include Relational Databases, NoSQL Databases, and Big Data platforms:

- MySQL, and PostgreSQL are examples of Open Source Relational Database Management Systems (RDBMS), and IBM Db2 and SQL Server are examples of commercial RDBMSes and are also available as Cloud services.
- MongoDB and Apache Cassandra are examples of NoSQL databases.
- Apache Hadoop and Apache Spark are used for Big Data analytics.

Data Integration and Transformation Tools: include Apache Airflow and Apache Kafka.

Data Visualization Tools: include commercial offerings such as Cognos Analytics, Tableau and PowerBI and can be used for building dynamic and interactive dashboards.

Code Asset Management Tools: Git is an essential code asset management tool. GitHub is a popular web-based platform for storing and managing source code. Its features make it an ideal tool for collaborative software development, including version control, issue tracking, and project management.

Development Environments: Popular development environments for Data Science include Jupyter Notebooks and RStudio.

- Jupyter Notebooks provides an interactive environment for creating and sharing code, descriptive text, data visualizations, and other computational artifacts in a web-browser based interface.
- RStudio is an integrated development environment (IDE) designed specifically for working with the R programming language, which is a popular tool for statistical computing and data analysis.