**Data Scientist Training**

**What is Data Science?**

* Data science is the practice of extracting valuable insights from vast datasets to guide strategic decision-making.
* Data science careers offer diverse paths, often involving mathematics, programming, and a curiosity for data exploration.
* Successful data scientists exhibit qualities like curiosity, critical judgment, and an aptitude for constructive argumentation.
* The data science field is characterized by high demand, resulting in attractive remuneration for skilled professionals.
* A Data Scientist's daily routine can vary significantly depending on the project's nature.
* A wide array of algorithms is available for extracting insights from data.
* Big Data plays a pivotal role in driving digital transformation across industries.
* Cloud computing is a fundamental technology in modern data science.
* Data mining techniques are essential for uncovering patterns and knowledge from data.
* Tools like Hadoop, HDFS, Hive, and Spark are employed for processing Big Data.
* Deep learning, machine learning, and regression are critical data science topics.
* Data science applications span diverse domains, solving complex problems.
* Companies can harness data science to address age-old challenges with innovative solutions.
* Data science contributes significantly to saving lives and improving various aspects of society.
* Careers in data science offer exciting opportunities, with mathematics and statistics being essential foundations.
* Reports in data science adhere to specific structures, and career roadmaps provide guidance.
* Case studies and projects offered practical application of the knowledge acquired during the course.

**Deep Learning and Machine Learning**

* Big Data has five characteristics: velocity, volume, variety, veracity, and value.
* The five cloud computing characteristics are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.
* Data mining has a six-step process: goal setting, selecting data sources, preprocessing, transforming, mining, and evaluation.
* The availability of so many disparate amounts of data created by people, tools, and machines requires new, innovative, and scalable technology to drive transformation.
* Deep learning utilizes neural networks to teach itself patterns in inputs and outputs. Machine learning is a subset of AI that uses computer algorithms to learn about data and make predictions without explicitly programming the analysis methods into the system.
* Regression identifies the strength and amount of the correlation between one or more inputs and an output.
* Skills involved in processing Big Data include the application of statistics, machine learning models, and some computer programming.
* Generative AI, a subset of artificial intelligence, focuses on producing new data rather than just analyzing existing data. It allows machines to create content, including images, music, language, computer code, and more, mimicking creations by people.

**Careers and Recruiting in Data Science**

* Data Science helps physicians provide the best treatment for their patients, helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
* Companies can start on their data science journey by capturing data. Once they have data, they can begin analyzing it.
* Everyone who uses the Internet generates mass amounts of data daily.
* Amazon and Netflix use recommendation engines, and UPS uses data from customers, drivers, and vehicles to use the drivers’ time and fuel efficiently.
* The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.
* The report should present a thorough analysis of the data and communicate the project findings.
* Companies should look for someone excited about working with the data in their particular industry. They should seek out someone curious who can ask interesting, meaningful questions about the types of data they intend to collect. They should hire people who love working with data, are fluent in statistics, and are competent in applying machine learning algorithms.
* A clearly organized and logical report should communicate the following to the reader:
  + What they gain by reading the report
  + Clearly defined goals
  + The significance of your contribution
  + Appropriate context by giving sufficient background
  + Why this work is practical and useful
  + Conjecture plausible future developments that might result from your work

**Data Science Tools Overview**

* 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.

**Open Source Tools For Data Science**

This reading provides a summary of key open-source tools for Data Science covered in the Part 1 and Part 2 videos of this course.

They are broadly classified as -

* **Data Management Tools** - Facilitates the storage, organization, and retrieval of data. Includes Relational Databases, NoSQL Databases, and Big Data platforms.
* **Data Integration and Transformation Tools** - Streamlines data pipelines and automate data processing workflows. Task of data integration and transformation in the classic data warehousing world is to Extract, Transform, and Load (ETL).
* **Data Visualization Tools**- Provides graphical representation of data and assist with communicating insights.
* **Model Deployment, Monitoring and Assessment Tools**- Supports the building, deploying, monitoring, and evaluation of data and machine learning models.
* **Data Asset Management Tools**- Organizes and manages data, enforce access controls, and ensure asset backups.
* **Code Development and Execution Tools** - ProvideS environments for developing, testing, and deploying code, offering computational resources to execute it.
* **Code Asset Management Tools** - Enables the storage and management of code, track changes, and support collaborative development.

**Data Management Tools**

**MySQL**

* Popular open source **relational database management system** (RDBMS)
* Uses structured query language (SQL) to manage and store data.
* Common uses:
  + Web applications
  + Data warehousing
  + E-commerce

**PostgreSQL**

* Powerful and open source **relational database management system** (RDBMS)
* Emphasizes extensibility and SQL compliance.
* Offers advanced features such as:
  + Support for JSON
  + Full-text search
  + Spatial data

**Apache CouchDB**

* Document-oriented **NoSQL** database
* Uses JSON to store data
* Highly scalable
* Fault-tolerant
* Easy to use

**MongoDB**

* Document-oriented **NoSQL** database
* Stores data in a flexible JSON
* Provides:
  + Scalability
  + High availability
  + Data distribution
* Suitable for modern web applications that handle large volumes of unstructured data

**Apache Cassandra**

* Highly scalable, distributed Document-oriented **NoSQL** database
* Can handle large amounts of structured and unstructured data across many commodity servers.
* Offers:
  + High availability
  + Fault tolerance
  + Tunable consistency levels
* Suitable for mission-critical applications

**Hadoop Distributed File System (HDFS)**

* Designed to work with large datasets like Apache Hadoop in a distributed computing environment
* High-throughput data processing by splitting files into blocks (default 128MB), and these blocks are distributed across multiple DataNodes
* Data is replicated across different DataNodes ensuring high availability and fault tolerance
* Scalable and efficient

**Ceph**

* Free, open source software-defined storage platform suitable for hybrid cloud environments
* Designed for modern data centers
* Provides highly scalable, unified storage system that can be used for object storage (like AWS S3), block storage (like virtual disks for VMs), and file storage (like NFS) under one unified system
* High performance, availability and reliability

**Elasticsearch**

* Primarily a distributed RESTful search engine and analytics tool
* Based on the Lucene library.
* Full-text search, real-time data analytics
* Highly scalable
* Easy to use
* Powerful querying capabilities
* Real-time data indexing for fast document retrieval.

**Data Integration and Transformation Tools**

**Apache Airflow**

* Open-source platform for programmatically authoring, scheduling, and monitoring workflows
* Created originally by Airbnb
* Allows users to define and execute complex workflows
* Support for:
  + Task dependencies
  + Parallelism
  + Error handling

**Kubeflow**

* An open-source machine learning toolkit that allows execution of data science pipelines on top of Kubernetes.
* Provides a platform for building, deploying, and managing end-to-end machine learning workflows at scale
* Support for:
  + Distributed training
  + Model serving
  + Hyperparameter tuning

**Apache Kafka**

* Distributed streaming platform that allows applications to publish, process, and subscribe to streams of records in real-time
* Created originally from LinkedIn.
* It is scalable, fault-tolerant, and high-throughput
* Suitable for building mission-critical, data-intensive applications

**Apache NiFi**

* An open-source data integration platform that allows users to automate the flow of data between systems
* Provides a web-based user interface for designing and managing data flows
* Support for:
  + Data routing
  + Transformation
  + Enrichment
  + Among other capabilities

**Apache Spark SQL**

* A module in the Spark ecosystem that provides a programming interface for working with structured data using:
  + SQL
  + Data frames
  + Datasets
* Supports a wide range of data sources and provides optimized performance for complex data processing tasks.

**Node-RED**

* An open-source visual programming tool for wiring together hardware devices, APIs, and online services
* Allows users to create event-driven flows of messages
* low in resource consumption that it even runs on tiny devices like a Raspberry Pi.
* Support for:
  + Data transformation
  + Filtering
  + Aggregation

**Data Visualization Tools**

**PixieDust**

* Open-source library for creating interactive, exploratory data visualizations in Python and Jupyter notebooks
* Provides a range of built-in visualizations and data connectors
* Support for customization and extensibility through third-party libraries

**Hue**

* Open-source web interface for analyzing and visualizing large datasets in Apache Hadoop
* Offers a user-friendly experience for exploring data and creating visualizations
* No need for programming skills; can create visualizations from SQL queries

**Kibana**

* Open-source data visualization tool that allows users to interact with their data through a web-based interface
* Commonly used with Elasticsearch to analyze and visualize large datasets

**Apache Superset**

* A modern, enterprise-ready business intelligence web application that makes it easy to visualize and explore large datasets
* Offers a rich set of data visualization options, including:
  + Charts
  + Tables
  + Maps
  + Geospatial analysis
  + Real-time data processing

**Model Deployment Tools**

**Apache PredictionIO**

* Open-source machine learning server built on a scalable and distributed infrastructure
* Allows developers to quickly build, evaluate, and deploy predictive engines for various use cases such as:
  + Recommendation
  + Classification
  + Clustering

**Kubernetes**

* Open-source platform for container orchestration
* Automatically launches, scales, and manages containerized applications
* Offering features like:
  + Automatic scaling
  + Self-healing
  + Load balancing
* Enables the management and orchestration of containers across numerous hosts

**Apache Seldon**

* Open-source platform for deploying and managing machine learning models on Kubernetes
* Provides a way to:
  + Serve models at scale
  + Automate model deployment workflows
  + Monitor the performance of deployed models in real-time

**MLeap**

* Open-source library for serializing and deserializing learning models in a cross-platform file
* Gives users the ability to export models from different machine learning libraries and frameworks, such as:
  + Spark
  + Scikit-learn
  + TensorFlow
* Implements them in high-throughput, low-latency production environments

**TensorFlow Lite**

* Open-source tool for running machine learning models on mobile and embedded devices
* Allows effective inference on mobile and embedded platforms
* Supports a variety of hardware accelerators such as:
  + CPUs
  + GPUs
  + Custom ASICs

**Red Hat OpenShift**

* Container application framework based on Kubernetes
* With characteristics like automation, scalability, and security
* Offers a method for creating, deploying, and managing containerized applications

**TensorFlow Serving**

* Open-source utility that serves machine learning models in real-world settings
* Supports both HTTP and gRPC interfaces for serving predictions
* Provides high scalability and low latency deployment and management of TensorFlow models

**TensorFlow.js**

* Open-source library for building and deploying machine learning models in JavaScript
* Allows you to train and execute models directly in the browser or on Node.js
* Supports a wide range of model architectures, including neural networks, decision trees, and k-nearest neighbors

**Model Monitoring and Assessment Tools**

**ModelDB**

* Open-source platform for managing machine learning models and experiments
* Provides a way to track and reproduce experiments, version models, and collaborate with team members

**Prometheus**

* Freely available monitoring system that collects and stores metrics in real-time from different sources
* Allows you to visualize and set alerts on the health and performance of systems and apps
* Supports a variety of data gathering methods, such as HTTP endpoints, exporters, and agents

**IBM AI Fairness 360**

* Open-source toolkit for detecting and mitigating bias in machine learning models
* Provides a way to measure the fairness and bias of models, as well as a set of algorithms for mitigating bias and creating fairer models

**IBM AI Explainability 360**

* Open-source toolkit for explaining the behavior and decisions of machine learning models
* Provides a way to measure the explainability and interpretability of models, as well as a set of algorithms for generating explanations and visualizations of model behavior

**IBM Adversarial Robustness 360 Toolbox**

* Free and open-source library for protecting machine learning models from adversarial attacks
* Includes a method for measuring model robustness and vulnerability
* Includes a set of algorithms for improving model robustness and detecting adversarial examples

**Code Development and Execution Tools**

**Jupyter IDE**

* Open-source effort
* Supports:
  + Julia
  + Python
  + R development with Jupyter Notebook
  + JupyterLab
  + JupyterHub
* Create and share documents containing:
  + Live code
  + Equations
  + Visualizations
  + Narrative text
* JupyterLab includes:
  + Customized notebook organization
* JupyterHub extends all these capabilities to the enterprise

**RStudio**

* For developers
* Free and open-source IDE
* Built to manage and execute R code
* Works on all platforms
* Includes:
  + Version control
  + Project management capabilities

**Microsoft Visual Studio**

* An IDE that supports a variety of programming languages, including:
  + C
  + C++
  + C++/CLI
  + Visual Basic.NET
  + C#
  + F#
  + JavaScript
  + TypeScript
  + XML
  + XSLT
  + HTML
  + CSS
* Using plug-ins, supports:
  + Python
  + Ruby
  + Node.js
  + M
  + Other languages

**PyCharm**

* Primarily a subscription-based IDE environment
* Offers 16+ additional tools for coding assistance, testing, and web development
* Supports scientific development with IPython integration and Matplotlib and NumPy support
* Also offers a free community-based, open-source IDE with limited capabilities

**Spyder**

* Free, open-source Python-based IDE designed by and for scientists, engineers, and data analysts
* Features a unique combination of comprehensive development tools for:
  + Advanced editing
  + Analysis
  + Debugging
  + Profiling
  + Visualization capabilities

**Anaconda Navigator**

* Open-source GUI-based Navigator that supports Python development and integrates with:
  + Eclipse and PyDev
  + IDLE
  + IntelliJ
  + Microsoft Visual Studio Code (VS Code)
  + Ninja IDE
  + PyCharm
  + Python for Visual Studio Code
  + Python Tools for Visual Studio (PTVS)
  + Spyder
  + Sublime Text
  + Wing IDE

**Code Asset Management Tools**

**Git**

* Open-source version control system for tracking changes in code and collaboration among developers
* Provides a way to manage and organize code changes, collaborate on code development, and maintain a history of code revisions

**GitLab**

* Web-based Git repository manager
* Provides a complete DevOps platform for:
  + Source code management
  + Continuous integration and deployment
  + Monitoring
* Enables teams to collaborate on:
  + Code development
  + Automate build and deployment processes
  + Track metrics and performance across the entire software development lifecycle

**GitHub**

* Web-based Git repository hosting service that provides a platform for developers to collaborate on code and manage software projects
* Enables users to:
  + Create, fork, and contribute to open source projects
  + Track changes in code
  + Manage issues
  + Pull requests

**Bitbucket from Atlassian**

* Web-based Git repository hosting service
* Provides a platform for developers to collaborate on code and manage software projects, with features like:
  + Pull requests
  + Code review
  + Branch permissions

**Languages of Data Science**

* You should select a language to learn depending on your needs, the problems you are trying to solve, and whom you are solving them for.
* The popular languages are Python, R, SQL, Scala, Java, C++, and Julia.
* For data science, you can use Python's scientific computing libraries like Pandas, NumPy, SciPy, and Matplotlib.
* Python can also be used for Natural Language Processing (NLP) using the Natural Language Toolkit (NLTK).
* Python is open source, and R is free software.
* R language’s array-oriented syntax makes it easier to translate from math to code for learners with no or minimal programming background.
* SQL is different from other software development languages because it is a non-procedural language.
* SQL was designed for managing data in relational databases.
* If you learn SQL and use it with one database, you can apply your SQL knowledge with many other databases easily.
* Data science tools built with Java include Weka, Java-ML, Apache MLlib, and Deeplearning4.
* For data science, popular program built with Scala is Apache Spark which includes Shark, MLlib, GraphX, and Spark Streaming.
* Programs built for Data Science with JavaScript include TensorFlow.js and R-js.
* One great application of Julia for Data Science is JuliaDB.

**Libraries, APIs, Datasets and Models**

* Python offers a diverse library ecosystem for data science, covering scientific computing (Pandas, NumPy), visualization (Matplotlib, Seaborn), and high-level machine learning (Scikit-learn). These libraries offer tools for data manipulation, mathematical operations, and simplified machine learning model development.
* Application Programming Interfaces (APIs) facilitate communication between software components. REST APIs, specifically, facilitate internet communication and access resources like storage. Key API terms include client (user or code accessing it), resource (service or data), and endpoint (API's URL).
* Machine learning models analyze data and identify patterns to make predictions and automate complex tasks—the three fundamental types of machine learning are supervised, unsupervised, and reinforcement learning. Supervised learning includes regression and classification models for predictive modeling and pattern recognition. Deep learning, an advanced subset of machine learning, mimics the brain's processing, enabling intricate problem-solving in various domains.
* The Community Data License Agreement (CDLA) facilitates open data sharing by providing clear licensing terms for distribution and use, and the IBM Data Asset eXchange (DAX) site contains high-quality open data sets.
* The Model Asset eXchange (MAX) provides a wealth of pre-trained deep learning models, empowering developers with readily deployable solutions for various business challenges.

**Additional Sources of Datasets**

In this reading, you will learn about:

* Open datasets and sources
* Proprietary datasets and sources
* Dataset license

**Open datasets and sources**

In this data-driven world, some datasets are freely available for anyone to access, use, modify, and share. These are called **open datasets**.  
Open datasets include a public license and are very useful for your journey as a Data Scientist. Some of the most informative open dataset sources are listed below.

**Government Data:**

* <https://www.data.gov/>
* <https://www.census.gov/data.html>
* <https://data.gov.uk/>
* <https://www.opendatanetwork.com/>
* <https://data.un.org/>

**Financial Data Sources:**

* <https://data.worldbank.org/>
* <https://www.globalfinancialdata.com/>
* <https://comtrade.un.org/>
* <https://www.nber.org/>
* <https://fred.stlouisfed.org/>

**Crime Data:**

* <https://www.fbi.gov/services/cjis/ucr>
* <https://www.icpsr.umich.edu/icpsrweb/content/NACJD/index.html>
* <https://www.drugabuse.gov/related-topics/trends-statistics>
* <https://www.unodc.org/unodc/en/data-and-analysis/>

**Health Data:**

* <https://www.who.int/gho/database/en/>
* <https://www.fda.gov/Food/default.htm>
* <https://seer.cancer.gov/faststats/selections.php?series=cancer>
* <https://www.opensciencedatacloud.org/>
* <https://pds.nasa.gov/>
* <https://earthdata.nasa.gov/>
* <https://www.sgim.org/communities/research/dataset-compendium/public-datasets-topic-grid>

**Academic and Business Data:**

* <https://scholar.google.com/>
* <https://nces.ed.gov/>
* <https://www.glassdoor.com/research/>
* <https://www.yelp.com/dataset>

**Other General Data:**

* <https://www.kaggle.com/datasets>
* <https://www.reddit.com/r/datasets/>

**Propriety datasets and sources**

Proprietary datasets contain data primarily owned and controlled by specific individuals or organizations. This data is limited in distribution because it is sold with a licensing agreement.  
Some data from private sources cannot be easily disclosed, like public data.

National security data, geological, geophysical, and biological data are examples of propriety data. Copyright laws or patents usually bind this type of data. Proprietary datasets that mainly contain sensitive information are less widely available than open datasets.

Some standard propriety dataset sources are listed below.

**Health Care:**

<https://www.sgim.org/communities/research/dataset-compendium/proprietary-datasets>

**Financial Market data:**

<https://datarade.ai/data-categories/proprietary-market-data>

**Google Cloud based datasets:**

<https://cloud.google.com/datasets>

**Dataset licenses**

When you select a dataset, it is necessary to look into the license. A license explains whether you can use that dataset or not; or explains if you have to accept certain guidelines to use that dataset. The different license types are listed below.

1. **PUBLIC DOMAIN MARK - PUBLIC DOMAIN**  
   When a dataset has a Public Domain license, all the rights to use, access, modify and share the dataset are open to everyone. Here there is technically no license.
2. **OPEN DATA COMMONS PUBLIC DOMAIN DEDICATION AND LICENSE – PDDL**  
   Open Data Commons license has the same features as the Public Domain license, but the difference is the PDDL license uses a licensing mechanism to give the rights to the dataset.
3. **CREATIVE COMMONS ATTRIBUTION 4.0 INTERNATIONAL CC-BY**  
   This license allows users to share and modify a dataset, but only if they give credit to the creator(s) of the dataset.
4. **COMMUNITY DATA LICENSE AGREEMENT – CDLA PERMISSIVE-2.0**  
   Like most open-source licenses, this license allows users to use, modify, adapt, and share the dataset, but only if a disclaimer of warranties and liability is also included.
5. **OPEN DATA COMMONS ATTRIBUTION LICENSE - ODC-BY**  
   This license allows users to share and adapt a dataset, but only if they give credit to the creator(s) of the dataset.
6. **CREATIVE COMMONS ATTRIBUTION-SHAREALIKE 4.0 INTERNATIONAL - CC-BY-SA**  
   This license allows users to use, share, and adapt a dataset, but only if they give credit to the dataset and show any changes or transformations, they made to the dataset. Users might not want to use this license because they have to share the work they did on the dataset.
7. **COMMUNITY DATA LICENSE AGREEMENT – CDLA-SHARING-1.0**  
   This license uses the principle of ‘copyleft’: users can use, modify, and adapt a dataset, but only if they don’t add license restrictions on the new work(s) they create with the dataset.
8. **OPEN DATA COMMONS OPEN DATABASE LICENSE - ODC-ODBL**  
   This license allows users to use, share, and adapt a dataset but only if they give credit to the dataset and show any changes or transformations they make to the dataset. Users might not want to use this license because they have to share the work they did on the dataset.
9. **CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL 4.0 INTERNATIONAL - CC BY-NC**  
   This license is a restrictive license. Users can share and adapt a dataset, provided they give credit to its creator(s) and ensure that the dataset is not used for any commercial purpose.
10. **CREATIVE COMMONS ATTRIBUTION-NO DERIVATIVES 4.0 INTERNATIONAL - CC BY-ND**  
    This license is also a restrictive license. Users can share a dataset if they give credit to its creator(s). This license does not allow additions, transformations, or changes to the dataset.
11. **CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-SHAREALIKE 4.0 INTERNATIONAL - CC BY-NC-SA**  
    This license allows users to share a dataset only if they give credit to its creator(s). Users can share additions, transformations, or changes to the dataset, but they cannot use the dataset for commercial purposes.
12. **CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 INTERNATIONAL - CC BY-NC-ND**  
    This license allows users to share a dataset only if they give credit to its creator(s). Users are not allowed to modify the dataset and are not allowed to use it for commercial purposes.

***Note: Additional license types exist. Any dataset you use will include details about its license.***

**Jupyter Notebooks and JupyterLab**

**Jupyter Notebooks on the Internet**

There are thousands of interesting Jupyter Notebooks available on the internet for you to learn from. One of the best sources is: <https://github.com/jupyter/jupyter/wiki>

It is important to notice that you can download such notebooks to your local computer or import them to a cloud based notebook tool so that you can rerun, modify, and apply what's explained in the notebook.

Very often, Jupyter Notebooks are already shared in a rendered view. This means that you can look at them as if they were running locally on your machine. But sometimes, folks only share a link to the Jupyter file (which you can make out by the \*.ipynb extension). In this case, you can pick the URL to that file and paste it to the NB-Viewer => <https://nbviewer.jupyter.org/>

The list of Jupyter Notebooks provides you with a huge collection of materials to explore. Therefore, it might be useful to give you some pointers to interesting notebooks. You have covered some examples with data in the labs. Let's highlight some useful data that further explores data science. In addition, as we have covered different tasks in data science, we will also provide a sample notebook for each task.

First, you start with exploratory data analysis, for which this notebook is highly recommended: <https://nbviewer.jupyter.org/github/Tanu-N-Prabhu/Python/blob/master/Exploratory_data_Analysis.ipynb>

For data integration/cleansing at a smaller scale, the python library\_pandas\_is often used. For this task, you can have a look at this notebook: <https://towardsdatascience.com/data-cleaning-with-python-using-pandas-library-c6f4a68ea8eb>

If you want to know more about clustering, have a look at this notebook: <https://nbviewer.jupyter.org/github/temporaer/tutorial_ml_gkbionics/blob/master/2%20-%20KMeans.ipynb>

And finally, if you want an in-depth notebook on the\_iris\_dataset, have a look at this: <https://www.kaggle.com/lalitharajesh/iris-dataset-exploratory-data-analysis>

**Module Summary**

* Jupyter Notebooks are used in Data Science for recording experiments and projects.
* Jupyter Lab is compatible with many files and Data Science languages.
* There are different ways to install and use Jupyter Notebooks.
* How to run, delete, and insert a code cell in Jupyter Notebooks.
* How to run multiple notebooks at the same time.
* How to present a notebook using a combination of Markdown and code cells.
* How to shut down your notebook sessions after you have completed your work on them.
* Jupyter implements a two-process model with a kernel and a client.
* The notebook server is responsible for saving and loading the notebooks.
* The kernel executes the cells of code contained in the Notebook.
* The Jupyter architecture uses the NB convert tool to convert files to other formats.
* The Anaconda Navigator GUI can launch multiple applications on a local device.
* Jupyter environments in the Anaconda Navigator include JupyterLab and VS Code.
* You can download Jupyter environments separately from the Anaconda Navigator, but they may not be configured properly.
* Additional open-source Jupyter environments include JupyterLab, JupyterLite, VS Code, and Google Colaboratory.
* JupyterLite is a browser-based tool.