Curriculum

MY COURSES

Data Science

2309FT Data Science [ALX] - Group A

2309FT Data Science [ALX] - Group A

6 Sprints 95 Lessons Sprint Feedback

Explore101

1 **Welcome to ExploreAI Academy**

In this lesson, we take a look at the ExploreAI Data Science course.

2 **An introduction to data science**

The world of data is vast and dynamic, constantly changing as our access to information increases. Understanding the fundamentals of data enables us to craft insights, make informed decisions, and unlock transformative opportunities. In this lesson, we'll embark on a journey to unravel the mysteries of data, big data, data-driven exploration and analysis, data science, and the diverse roles in the world of data. By the end of this lesson, you will know what data and big data are, how we approach data analysis and data science to derive meaningful insights and understand the in-demand attributes, skills, and tools of data professionals.

3 **Ways of work at ExploreAI**

The modern world requires us and our data solutions to be adaptable and efficient. As a result, to thrive in this environment, it's essential to understand and embrace effective ways of working and how this affects the planning, development, and maintenance of data solutions.

In this lesson, we'll take a look at various frameworks and methodologies that provide us with structured ways of working in teams, building complex data solutions, and measuring the value and impact of these solutions. We'll also take a look at some tools and best practices for the various phases of the product development lifecycle.

4 [Problem solving]

To ensure that we can solve problems of varying complexity and context, we need to understand how we solve problems. This helps us to solve problems because it gives us an intellectual toolset that can be applied to any problem and ensures that we correctly define problems.

In this lesson, we'll learn all about solving problems using various toolsets and methodologies. By the end of this lesson, you will be able to leverage problem-solving techniques such as 5W2H, MECE, and design thinking to approach and solve various problems including wicked problems.

5 [Flowcharts and pseudocode]

In order to break down a problem or task into smaller more manageable parts that we can solve in a systematic and logical way, we need programmatic thinking. In order to apply programmatic thinking to solving problems, we need a variety of tools, including algorithms, operators, flowcharts, pseudocode, and conditional statements.

In this lesson, we'll take a deep dive into using flowcharts, pseudocode, and conditional statements to solve a variety of logic problems.

6 [End of Module feedback]

Now that you have completed this module, please let us know how it went. Thanks in advance

Preparing Data

1 [Module overview: DS Preparing data]

In this module, we focus on developing essential data and spreadsheet skills, enabling us to effectively work with data to draw meaningful insights and make informed decisions. By exploring data in Google Sheets, we will learn how to retrieve, manipulate, analyse, and visualise data.

2 [Data sources and access]

Data plays a pivotal role in conducting effective analysis. Significant effort and attention needs to go into the approaches used to retrieve and access it.

In this lesson, we explore ways to source the right data and how to take accountability for it through data governance principles and ethics. We also look at spreadsheets, and more specifically Google Sheets, as a tool we can use to access our data.

3 [An introduction to using data]

To be able to draw meaningful insights from a spreadsheet dataset, we need to understand how to use data. This helps us transform our dataset from raw, unprocessed data to information that can be used to make decisions.

In this lesson, we will be introduced to using data in spreadsheets as we look into why it is necessary. By the end of this lesson, you

will be able to identify the different data types in spreadsheets, know when and how to make data visible, as well as know how to use a few functions in spreadsheets.

4 [Data aggregations and descriptive statistics]

Descriptive statistics is a way to describe and summarise the characteristics of the data you are working with.

In this lesson we will learn about the different kinds of descriptive statistics and how to calculate them. By the end of this lesson, you will be able to calculate central tendency and spread as well as use pivot tables to summarise data.

5 [An introduction to data visualisation]

Data visualisation is the art and science of visually representing data and complex information as charts, graphs, and other visual elements. It enables us to transform raw data into intuitive and accessible representations of data that we can use to identify patterns, trends, and relationships in the data.

In this lesson, we'll take a look at what data visualisation is and how to choose the most appropriate visualisation based on our data and the story we want to tell. By the end of this lesson, you will know how to choose appropriate visualisations, create them in Google Sheets, and be able to interpret and analyse them.

6 [Integrated project: Access to drinking water (Part 1)]

In this first part of the integrated project, we investigate access to safe and affordable drinking water. You will need to create features, summaries, and visualisations as per the instructions in order to complete the compulsory case study MCQ assessment.

7 [Data formatting]

In this lesson, we will learn about data formatting and discover why it is crucial to format data in order to guarantee and maintain data integrity. Additionally, we'll learn about the typical data type and structural issues we might encounter when dealing with spreadsheets.

8 [Data cleaning]

Data preparation efforts and approaches are crucial and significantly impact the outcomes of our analysis, sometimes, even more than the analysis methods themselves. Keen attention needs to be applied to the data cleaning process to ensure that only high-quality data is utilised for effective analysis. Moreover, it is equally important to maintain data quality consistently throughout its entire lifecycle.

In this lesson, we are introduced to the concepts of data cleaning and data integrity and the crucial role they play in enhancing and maintaining data quality. We will also see how we can implement these concepts practically in Google Sheets.

9 [Samples and distributions]

In this lesson, we will introduce the concept of probability and delve into how to apply the empirical probability approach to data. We will explore the concept of sample and population data and how sample sizes affect confidence intervals and the margin of error.

10 [Spreadsheet functions]

To harness the power of data, we need to master the art of data manipulation and analysis through the use of spreadsheet functions. These toolsets empower us to solve problems and extract valuable information from complex datasets efficiently and accurately.

In this lesson, we'll cover a wide range of spreadsheet functions. By the end of this lesson, you will be able to employ techniques such as VLOOKUP, regular expressions, and logical functions to sort, filter, and format data to make it more accessible and meaningful.

11 [Integrated project: Access to drinking water (Part 2)]

In this second part of the integrated project, we extend our investigation on access to safe and affordable drinking water. You will need to create features, summaries, and visualisations as per the instructions in order to complete the compulsory case study MCQ assessment.

12 [Identifying patterns]

To ensure that we can effectively identify data patterns, we need to understand the underlying concepts and techniques. This understanding helps us to identify trends, establish relationships, and make data-driven decisions.

In this lesson, we'll explore the line of best fit, correlation coefficients, R-squared values and heatmaps. By the end of this lesson, we will be able to use these tools to interpret relationships, identify data patterns and make predictions.

13 [Accuracy]

Understanding model accuracy equips us with the ability to create reliable models that can be applied to various problems.

In this lesson, we'll learn about model accuracy metrics and how to use these to assess our models. We'll uncover common challenges to model accuracy, such as bias, variance, overfitting, and underfitting, and explore strategies to address these obstacles. Finally, we will learn how to use polynomial trend lines to find the best fit for our data. By the end of this lesson, you will be able to assess the accuracy of models and apply techniques to improve the fit of models to our data.

14 [Drawing and testing assumptions]

Hypothesis testing serves as the framework through which we can draw and examine assumptions about our population data or the relationships between variables in a dataset. These assumptions help us understand the characteristics of our data and provide insights into the reliability and generalisability of the results.

In this lesson, we will explore the fundamental concepts of hypothesis testing and its importance in statistical analysis. By the end of this lesson, you will be able to formulate hypotheses and test their validity using various statistical tests.

15 [End of Module feedback]

Now that you have completed this module, please let us know how it went. Thanks in advance

SQL

1 [Module overview: SQL]

In this module, we delve into the world of SQL (Structured Query Language) to equip us with the necessary skills for managing and extracting valuable insights from data. SQL is a powerful tool used for data manipulation and retrieval in relational databases, making it a fundamental language for data professionals.

Throughout this module, we will explore the core principles of SQL and learn how to query relational databases efficiently, perform data transformations, and gain proficiency in designing complex queries.

2 [Database concepts]

Databases play a critical role in enabling us to extract meaningful insight from data because they provide a structured way to store, manage, and retrieve large amounts of data.

In this lesson, we take a look at fundamental database concepts and explore how we can interact with databases using a Database Management System (DBMS).

3 [SQL basics]

Understanding the fundamentals of SQL and the SQL sublanguages is crucial for effectively managing and manipulating data.

In this lesson, we'll introduce the five principal SQL sublanguages and their respective commands. We will focus on DDL commands to learn how to define data structure and DML commands to learn to manipulate data. Additionally, we'll learn how to use the WHERE clause and AND operator to modify specific records based on conditions. By the end of this lesson, you'll have the knowledge and skills to navigate SQL databases and apply techniques to manage and manipulate data.

4 [Querying with SQL]

Data Query Language (DQL) statements in SQL allow us to search and retrieve data from one or more database tables, using various conditions, criteria, and filters.

In this lesson, we take a deep dive into querying with SQL to search and retrieve data from a database using various SQL keywords.

5 [Integrated project: Beginning our data-driven journey in Maji Ndogo]

In this first part of the integrated project, we dive into Maji Ndogo's expansive database containing 60,000 records spread across various tables. As we navigate this trove of data, we'll use basic queries to familiarise ourselves with the content of each table. Along the way, we'll also refine some data using DML.

6 [SQL in production]

SQL is integral to the success of modern data-driven systems, facilitating everything from efficient data storage and retrieval to supporting complex analytics and reporting. Due to SQL's flexibility and wide-ranging functionality, it allows for diverse applications in production environments used by organisations day to day.

In this lesson, we take a look at how organisations use SQL in production environments and how technology stacks, application architectures, and business requirements influence how we use SQL. We also take a look at how we can write SQL queries in Python notebooks.

7 [Querying in notebooks]

Jupyter notebooks are a powerful tool that combines code, documentation, and visualisation, making it an ideal platform for data analysis and database querying tasks.

In this lesson, we'll look at using Jupyter notebooks to interact with SQL and SQL databases. Additionally, we demonstrate how to use SQL best practices and execute basic SQL queries in Jupyter notebooks.

8 [SQL: Numeric functions and aggregations]

In this lesson, we dive into the various types of SQL functions used to perform calculations on data and manipulate the data within the database to produce aggregate summaries and calculations.

9 [SQL: Window functions]

SQL window functions, also known as analytic functions, are a powerful way to analyse and manipulate data within SQL queries by allowing us to perform calculations that consider the context of surrounding rows.

In this lesson, we are introduced to SQL window functions and how

to harness their capabilities to perform advanced data analysis tasks. We will learn how to rank data, calculate running totals, and access values from neighbouring rows.

10 **SQL: String, date, and miscellaneous functions**

Mastering the complexities of string, date, and miscellaneous functions is paramount for effective data manipulation, as they serve as fundamental tools for transforming and analysing data with precision.

In this lesson, we'll take a look at manipulating textual and DateTime data in SQL. Additionally, we will explore miscellaneous functions that are valuable for converting between various data types.

11 **SQL: Control flow functions**

Understanding control flow functions is crucial for enabling conditional logic within queries, allowing for dynamic data retrieval and manipulation based on specific criteria.

In this lesson, we'll introduce two main control flow functions, the CASE and IF statements. We will focus on the syntax and practical applications of these functions to learn how to categorise and conditionally manipulate data. Additionally, we'll learn how to nest these functions and combine them with other SQL features, such as aggregate functions and the GROUP BY clause. By the end of this lesson, you'll have the knowledge and skills to navigate control flow functions and apply techniques to handle data based on specific conditions.

12 **Integrated project: Clustering data to unveil Maji Ndogo's water crisis**

In this second part of the integrated project, we gear up for a deep analytical dive into Maji Ndogo's water scenario. Harness the power of a wide range of functions, including intricate window functions, to tease out insights from the data tables.

13 **SQL: Entity-relationship data models**

In this lesson, we dive into the various types of relationships that exist within a database and the Entity-Relationship Diagrams that can be designed based on these relationships. We dive deeper into the keys that are used to implement these relationships and how to create them practically in SQL.

14 [SQL: Joins and set operations]

In the world of relational databases, the ability to combine and manipulate data from multiple tables is essential. This lesson delves into the powerful concepts of SQL joins and unions while drawing parallels with set theory.

By the end of this lesson, you will have a comprehensive understanding of how to seamlessly integrate data from different tables and perform complex operations using these techniques.

15 [SQL: Optimising queries]

SQL in production operates on datasets that can contain hundreds of millions of rows of data. Poorly optimised SQL queries can use up an enormous amount of resources if they are applied incorrectly, and are difficult to read and modify.

In this lesson, we discuss advanced topics in SQL. We discuss how subqueries and Common Table Expressions enable more advanced analysis and how to optimise queries so that they execute quickly, are easy to read and understand, and are simple to modify as the needs arise.

16 [SQL: Views]

A fundamental component of relational database management systems is SQL views. SQL views are a powerful tool for simplifying complex queries, enhancing data security, and improving query performance.

In this lesson, we get to learn how views serve as virtual tables that improve data accessibility and enhance data security.

17 [SQL: Normalisation]

Data normalisation is the process of breaking down complex, unorganised datasets into a series of related tables that adhere to specific rules known as normal forms. By applying normalisation rules, data is organised in a structured manner, reducing data duplication and dependency issues.

In this lesson, we look at how we can apply database normalisation principles, known as normal forms, to design well-structured and efficient databases that promote data integrity and reduce redundancy.

18 [An introduction to NoSQL]

In this lesson, we will examine the key aspects of NoSQL databases. We will also learn about OLAP, OLTP, and NewSQL, as well as their applications and how they differ from standard relational databases.

19 [Integrated project: Weaving the data threads of Maji Ndogo's narrative]

In this third part of the integrated project, we will pull data from many different tables and apply some statistical analyses to examine the consequences of an audit report that cross-references a random sample of records.

20 [Integrated project: Charting the course for Maji Ndogo's water future]

In this final part of the project, we finalise our data analysis using the full suite of SQL tools. We will gain our final insights, use these to classify water sources, and prepare relevant data for our engineering teams.

21 [SQL exam: The Movie Database]

In this exam, you will be exploring the The Movie Database - an online movie and TV show database, which houses some of the most popular movies and TV shows at your fingertips. The TMDb database supports 39 official languages used in over 180 countries daily and dates all the way back to 2008.

22 [End of Module feedback]

Now that you have completed this module, please let us know how it went. Thanks in advance

Data Visualisation and Storytelling

1 [Module overview: DS Visualising data]

In this module, we explore data storytelling, communication, design, visualisation, dashboards and reports. Throughout this course, we will use Microsoft Power BI to build data models, create new features, and craft interactive dashboards and reports that will enable us to convey insights, provide actionable recommendations, foster collaboration, influence stakeholders, engage others in the data process, and build trust.

2 [Communicating our findings]

Effective communication is essential for data professionals to bridge the gap between technical analysis and interpretation, and decision-making. It enables us to convey insights, provide actionable recommendations, foster collaboration, influence stakeholders, engage others in the data process, and build trust. By mastering the art of visual storytelling and communication design, we can maximise the impact of our work and drive decision-making within organisations.

In this lesson, we will delve into the fundamentals of storytelling and impactful communication. By the end of this lesson, you will be able to leverage these skills to tell masterful data stories and increase the impact of your insights.

3 [Design for impactful communication]

Effective communication through design is a transformative skill for data professionals. It elevates raw data to the level of actionable insights, influences stakeholders by making the data relatable and persuasive, and ultimately drives decision-making by presenting information in a way that empowers organisations to make better choices.

In this lesson, we will explore key elements to enhance our design skills in order to create compelling data visualisations and presentations that will make a lasting impact.

4 [An introduction to dashboards and reports]

Dashboards and reports are powerful tools that enable data professionals to transform complex data into concise visual representations, providing quick and actionable insights.

In this lesson, we will explore the fundamentals of dashboards and reports, focusing on the features and functionalities of Microsoft Power BI as a leading dashboarding and reporting tool. We will take a look at various methods of importing and connecting data to Power BI and delve into sharing insights and deploying dashboards and reports using Power BI. By the end of this lesson, you will be able to navigate Power BI and connect to various data sources.

5 [Creating visuals in Power BI]

Creating compelling and informative visuals is a crucial aspect of data analysis and reporting. Data visualisations enable us to transform raw data into intuitive and accessible representations of data that we can use to identify patterns, trends, and relationships in the data.

In this lesson, we'll take a look at creating visualisations using Microsoft Power BI. We will explore the creation of various types of visuals, including line, column, 100% stacked column, 100% stacked area, scatter, and bubble charts. We'll also look at how the data structure can influence how we create visualisations in Power BI and how we can leverage filters on visualisations. By the end of this lesson, you will be able to create various meaningful visualisations in Power BI.

It is recommended that you revisit the lesson "An introduction to data visualisation" where we explored choosing appropriate visualisations based on our data and the story we want to tell.

6 [Integrated project: Visualising Maji Ndogo's past]

In this first part of the integrated project, we are introduced to updated data concerning the gender composition of queues at shared water taps in Maji Ndogo, and some new crime-related data.

7 [Formatting visuals in Power BI]

Appropriate formatting of data visualisations unlocks the data story we are trying to tell by enabling clarity and consistency in our visualisations. This makes it easier for the audience to interpret and draw meaningful insights from the visualisations presented.

In this lesson, we explore the formatting of visuals in Microsoft Power BI, including modifying charts, axes, and value labels, numerical precisions, fonts, colours, lines, and tooltips. By the end of this lesson, you will be able to create impactful and consistent visualisations in Power BI.

8 [Data models in Power BI]

Data models form the foundation of meaningful insights and interactive reports in Power BI. Much like SQL (Structured Query Language) serves as the foundation for database management, Power BI relies on data models to seamlessly connect, organise, and transform data into actionable insights. In Power BI we have a structured representation of our data, with tables, relationships, and calculations.

In this lesson, we take a look at viewing, creating, and managing data models in Power BI, and explore how our data model and data granularity impact our reports. By the end of this lesson, you will be able to create and manage data models in Power BI, know how to resolve model errors and understand how data models impact our reports and dashboards.

9 [Data transformations in Power BI]

Data transformations are a crucial step in the data analysis process, enabling us to shape and refine our data. While Power BI is widely recognised for its prowess in visualisations, reporting, and dashboarding, it also offers a robust suite of tools and functionalities that allow us to perform relatively complex transformations on our data, unlocking the true potential of our datasets for deeper insights and informed decision-making.

In this lesson, we'll take a look at how we can implement various data transformations in Power BI using Power Query Editor, including cleaning data, checking its integrity, and restructuring data to facilitate specific data stories in our reporting. By the end of this lesson, you will be able to apply basic data transformations in Power BI.

10 [Integrated project: Moulding data into visual stories in Maji Ndogo]

In this second part of the integrated project, we focus on data models. We'll import tables separately, clean data, and set up a working relational data model in PowerBI. We will also recreate our visuals with the new data model, and see how the new model affected our visuals. Our goal is to refine the visuals, customising text, colours, and fonts to make the visuals clear and simple.

11 [Calculated columns with DAX]

The ability to manipulate and interpret data using calculated columns is essential for crafting insightful and detailed reports.

In this lesson, we learn how to use DAX to create calculated columns. We delve into the syntax and functionalities of DAX for string concatenation, data readability enhancement, and dynamic data evaluation. We will explore how to effectively leverage DAX variables, functions, and operators to construct sophisticated data models and perform nuanced analysis. By the end of this lesson, we will be equipped with the skills to utilise DAX for advanced data manipulation, leading to more refined and impactful data visualisation in Power BI.

12 [DAX aggregations]

In the dynamic landscape of Power BI, mastering the art of data aggregation and transformation is pivotal for converting raw datasets into insightful reports.

In this lesson, we hone in on the strategic use of DAX aggregations to enhance data representation and create new tables. We will delve into the intricacies of leveraging DAX functions to aggregate and transform data, empowering us to craft purposeful and optimised data models for robust reporting and analysis in Power BI.

13 [Building reports and dashboards]

The creation of reports and dashboards in Power BI encapsulates the art of storytelling, design principles and practices, and leveraging the capabilities of data manipulation and visualisation tools within Power BI. It merges the technical aspects of data wrangling, visual presentation, and the finesse of storytelling to engage stakeholders, involve others in the data exploration process, and establish trust in the insights derived from data.

In this lesson, we'll see how we can use the various functionalities of Power BI together to create impactful, intuitive, and interactive reports and dashboards. By the end of this lesson, you will be able to create reports and dashboards using a variety of visualisations, filers, slicers, drills, buttons, and bookmarks, to engage and convince stakeholders.

14 [Exploratory Data Analysis in Power BI]

Exploratory Data Analysis (EDA) is a critical step in the data analysis process that involves examining and visualising data to understand their main characteristics, uncover patterns, and identify potential relationships between variables. The primary goal of EDA is to gain insights into data, generate hypotheses, and guide further analysis. By combining the capabilities of Power BI with the principles of EDA, analysts and decision-makers can efficiently explore and derive meaningful insights from their data.

In this lesson, we'll explore the ways in which we can use Microsoft

Power BI to unveil insights and patterns hidden within data. By the end of this lesson, you will be able to use various Power BI functions and features to apply EDA to gain insights into data.

15 **Integrated project: Communicating our findings in Maji Ndogo**

In this third part of the integrated project, we finalise our national survey report. We will use DAX to create measures and columns to enrich our data to ensure accurate and useful data representation on the dashboard. We put together all we have learned in the module to create the survey report.

16 **Integrated project: Transparency in tracking Maji Ndogo's water funds**

In this final part of the project, we use all the skills acquired in the course to build a public dashboard. Our mission is to communicate with transparency: Where did the money go? We will track the total budget against project completion, monitor teams' performance, and compare budgeted versus actual costs to flag potential corruption, promoting transparency and accountability in addressing Maji Ndogo's water crisis.

17 **End of Module feedback**

Now that you have completed this module, please let us know how it went. Thanks in advance

Python

1 **Module overview: Python**

In this module, we explore the versatile realm of Python programming, an essential language for various applications, from web development to data science. Python's readability and extensive libraries make it a powerful tool for both beginners and experienced developers. Throughout this course, we will delve into the core principles of Python, covering topics such as syntax, data structures, and control flow. By the end, you'll be equipped with the skills to write efficient and scalable Python code for a wide range of purposes.

2 **An introduction to Python**

Understanding Python's role and capabilities is essential to harnessing its power for complex data analysis and modelling.

In this lesson, we introduce Python's features, evolution, and real-world applications. We'll cover setting up Python tools, managing notebooks in environments like Jupyter Notebook and Google Colab, and submitting code on Athena. We'll also learn basic Python programming, including arithmetic operations and using the print() function.

By the end of this lesson, we'll be skilled in Python's fundamentals and ready for more advanced coding challenges.

3 [Python variables and data types]

Variables and data types form the basis of programming in Python. In this lesson, we'll introduce variables and explore their use. We'll also define the different data types and the operators that can be used to manipulate them.

By the end of this lesson, you should be comfortable using and manipulating different data types stored as variables.

4 [Python data structures]

In this comprehensive module on Python data structures, we'll embark on an exciting journey to understand and utilise Python's core data structures - tuples, lists, sets, and dictionaries. As we delve into each structure's theory and practice, our primary focus will be on applying these concepts to a captivating real-world task: constructing a digital twin of a farm. This practical application will not only illustrate the strengths and uses of each data structure but also provide you with hands-on experience in digital representation and simulation.

By the end of this module, you'll have developed a solid foundation in Python's data structures, equipped to represent complex real-world entities, like a farm, digitally. Expect to gain skills in data organisation, efficient information retrieval, and the ability to apply these structures creatively to solve practical problems in data science and beyond.

5 [Logic and loops in Python]

Understanding Python's logic and loops is key to developing efficient and powerful programming solutions. In this lesson, we explore how to use conditional statements like if, elif, and else in Python for decision-making. We'll also learn about VSCode and how to use it for Python coding. We'll introduce the basics of for and while loops, important for handling repetitive tasks and data, and learn how to control these loops. We'll also cover list and dictionary comprehensions. By the end of this lesson, you'll have a good grasp of Python's conditional statements and loops, ready to solve more complex problems.

6 [An introduction to version control]

Data science and data engineering projects often require multiple people to contribute code to a project simultaneously, or in short sequence. A remote repository for code can alleviate a lot of the complexity that comes with such projects, as it provides a way to ensure version control, collaboration and transparency in the project's history.

In this lesson, we'll learn how to use version control with GitHub.

7 [Python functions]

Python functions are pivotal in enhancing code modularity and reusability. They allow developers to encapsulate specific functionality, making code more organised and efficient.

In this lesson, we'll delve into the fundamental role of functions in Python programming, exploring their diverse applications. Beyond

mere understanding, we'll actively engage in creating our own functions, unveiling the versatile power they bring to the table. Through practical examples and hands-on exercises, we'll gain a comprehensive insight into the inner workings of Python functions, empowering us to write more maintainable and scalable code.

8 [Algorithms and algorithmic complexity]

It is important to measure the performance of algorithms and understand their scalability in order to optimise code, enhance efficiency, and ensure robust solutions that can handle increasing data sizes without compromising speed or functionality.

In this lesson, we'll explore the fundamental concepts of computational complexity and Big O notation, providing a clear understanding of algorithm efficiency. Delving into search algorithms, we'll examine their complexities and discuss the significance of efficient search strategies. We will also explore how to craft pseudocode for both search and sort algorithms, fostering a practical understanding of algorithmic design and analysis.

9 [Recursive and Lambda functions]

Understanding recursive and Lambda functions unlocks powerful and efficient coding strategies in Python. In this lesson, we explore recursive functions, focusing on their ability to simplify complex problems. We'll also learn about sorting algorithms like merge sort, demonstrating recursion's effectiveness. Finally, we look at Lambda functions and their role in writing more concise code, as well as how they can be used with Python's built-in functions. By the end of this lesson, you'll be adept at leveraging the strengths of recursive and Lambda functions for more streamlined and powerful Python programming.

10 [Classes, objects, and methods]

Encapsulating data and functionalities within classes can create a structured and efficient approach to coding. As a data analyst, it is fundamental to learn how to define classes, instantiate them, and use their methods to perform data analyses and other data manipulations which would result in reusability and abstraction in programming.

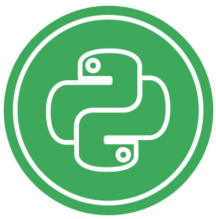
In this lesson, we delve into the fundamentals of object-oriented programming (OOP) by exploring Python classes.

11 [PEP 8]

In this lesson on PEP 8, we will navigate the world of Python coding standards. This journey is more than just a walkthrough of rules; it's an exploration into the ethos of writing clean, readable, and professional Python code. We'll dissect the core principles of PEP 8, delving into why these guidelines exist and how they enhance the quality of Python programming.

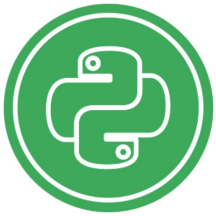
12 [An introduction to packages in Python]

In this lesson, we explore the concepts of modules and packages. Modules are smaller pieces of code stored in a single file, which can be called or imported to another Python program when needed. By doing so, we can structure our code in such a way that we can reuse code easily and efficiently. If we have a collection of modules on related topics, we can create a package, referring to the files in a specific directory, and import it into another script. This allows us to build up a library of code for future use.



13 [Data handling in Python]

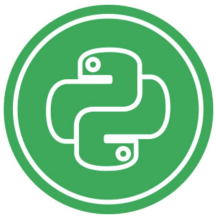
NumPy is a Python package that allows us to work with data and perform operations like the loading, analysing and storing of data. It provides high-performance, multidimensional array objects, numerical computing tools, and is fundamental in scientific computing. It also forms the basis of the Pandas library. In this lesson, we'll learn how to create, manipulate, and analyse data stored in NumPy arrays.



14 [Pandas for Data Science]

In this lesson, we delve deeper into using Pandas for manipulating and utilising DataFrame objects in Python. Given the frequent use of DataFrames in Data Science projects, we explore a combination of techniques to wrangle and process data.

We will learn how to effectively sort, filter, and transform data, as well as how to perform crucial operations like creating and deleting columns in a DataFrame. By the end of this lesson, we'll have a solid foundation in leveraging Pandas for commonly performed data tasks.



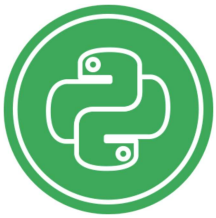
15 [Integrated Project: Understanding Maji Ndogo's agriculture Part 1]

This lesson is centred on the integrated project, applying skills learned in Pandas to understand the agricultural landscape of Maji Ndogo.

16 [Advanced string manipulation]

Regular expressions (regex) in Python are essential tools for anyone working with text data, enabling efficient searching, editing, and data manipulation. Mastering regex opens up a world of possibilities for data scientists and engineers, allowing for precise pattern matching and text processing in large datasets.

In this lesson, we'll explore the basics of Python's regex library, focusing on practical applications such as extracting specific information and using compiled regex objects for performance. We'll gain a solid understanding of regex patterns, methods, and their use in real-world data processing challenges, enhancing our data manipulation skills in Python, especially when combined with Pandas for complex data tasks.



17 [Python visualisations]

Harnessing the power of data visualisation in Pandas is instrumental for efficient data analysis and interpretation.

In this lesson, we will navigate the core principles of data visualisation using Python's Matplotlib and Pandas libraries. Starting with foundational tools like bar graphs and pie charts, we will explore their significance in representing categorical and proportional data. Subsequently, we will delve into line graphs and scatter plots, mastering the art of coding time series representations and identifying variable relationships. Through the integration of Seaborn and Matplotlib, we will enable the creation of sophisticated visualisations and interactive techniques, ensuring a well-rounded proficiency in data visualisation with Pandas.

18 [Integrated Project: Understanding and trusting data (Part 2)]

In this lesson, we're going to dive into the agricultural dataset to find patterns and try to get insights. We're also stopping for a moment to make sure our data is true and sound by validating it against other data sources.
This lesson consists of: 1. An Integrated Project notebook in which we're challenged to complete an analysis of the Maji Ndogo agricultural dataset. 2. An MCQ test related to the notebook.

19 [Statistics with Python]

Statistics is essential for analysing and interpreting data in projects, and Python's rich libraries make it a powerful tool for this purpose. This enables improved analysis, teamwork, and project transparency.

In this lesson, we'll learn how to use Python for Statistics. We'll start by covering some basic functionality of Python for the field of Statistics, before looking at the data science process as a whole, considering what we've learnt up to now.

20 [Statistics with Python [MCQ]]

Test your knowledge of statistics, sampling, distributions and hypothesis testing using Python.
Content to revise: All content in the Statistics with Python lesson. Since we're using Python to do statistics, packages like Pandas and Numpy.
Number of questions: 10

21 [Scripting and testing]

Effective software testing and debugging are paramount to ensuring the reliability and functionality of code, especially as projects increase in size. Python scripting serves as a powerful tool for developers and data scientists, enabling them to create efficient and reusable code that is functional across multiple platforms and environments.

In this lesson, we will delve into the critical role of software testing, exploring verification, validation, and the four levels of testing. We will then transition to error interpretation and debugging and explore some debugging techniques. We will also look at Python

scripting and include command line usage, script creation, and execution in different environments. Finally, we'll review how to integrate scripts back into Jupyter notebooks.

22 **More advanced Python**

Python is open-source in nature which means that users can contribute to its expansive ecosystem. It provides a collaborative environment that allows for the creation of specialised packages to enhance functionality and streamline complex tasks.

As a Data Scientist, we've encountered powerful packages like NumPy and Pandas, which are pivotal in data analysis. In this lesson, we will learn how to build our own package and share our solutions.

We will look at the steps involved in package creation, including writing code and setting up the necessary supporting files. We will also learn how to distribute our package online using GitHub and install it from anywhere.

23 **Integrated Project: Validating our data**

In this Code Challenge we're diving into the agricultural dataset again to continue to validate our data. Before we do that, we're pausing to build a data pipeline that will ingest and clean our data with the press of a button, cleaning up our code significantly. Once that's ready, we'll complete our data validation.

24 **Python exam**

In this exam, we test our understanding of Python and ability to apply our knowledge.

25 **End of Module feedback**

Now that you have completed this module, please let us know how it went. Thanks in advance

Regression

1 **Module overview: Regression**

In this module, we delve into the realm of regression analysis, covering essential techniques and methodologies crucial for predictive modelling and data-driven decision-making. We start by exploring fundamental concepts such as linear regression and model performance evaluation, gradually progressing to more advanced topics like multiple linear regression, variable selection, regularisation, ensemble methods, and bootstrapping.

2 **An introduction to machine learning**

There are certain key concepts that are often used in conjunction with solving data science problems. In this lesson, we are going to explain some of these methods and approaches to help form the foundation for more complex concepts in the weeks to come. We'll

first take a look at what is meant with machine learning, then look at predictive modelling, before examining the measures we can use to assess whether a model is performing well or not.

3 Linear models

Understanding linear models provides a foundation for regression analysis and predictive modelling. In this lesson, we delve into the basics of simple linear regression and its application in modelling the relationship between two variables to make predictions. We then look at the least squares method and how it is used to find the line of best fit. Finally, we learn how to implement a linear regression model using Python's scikit-learn library, evaluate its performance and interpret the results. By the end of this lesson, you'll possess the skills to apply simple linear regression effectively for insightful data analysis.

4 Model performance

Predictive models are useful tools that guide decision-making and forecast future trends. However, to ensure that our models are reliable and can deliver real-world value, we must thoroughly evaluate them to identify weaknesses and opportunities for improvement.

In this lesson, we will explore the various aspects of evaluating model performance such as the metrics and challenges that underpin this evaluation. We will also look at how to assess a model's ability to generalise to new data and why this is an important indicator of a model's real-world performance.

5 Multiple linear regression

Understanding multiple linear regression and its evaluation metrics is crucial, as it provides a powerful tool for uncovering relationships within data and making predictions. Multiple linear regression serves as the cornerstone for various advanced machine learning algorithms and statistical techniques, and is often used in fields such as finance, healthcare, marketing, and engineering.

In this lesson, we will cover the fundamentals of multiple linear regression, including its assumptions, implementation in Python using libraries like sklearn and statsmodels, and evaluation techniques. We'll explore how to check for linearity, multicollinearity, independence, homoscedasticity, normality, and outliers in regression models.

6 Variable selection and model persistence

Variables are the building blocks of machine learning models, and all contribute to and affect the model building process differently. Variable selection is necessary as it aims to improve model performance, speeding up model training, and reduce computational costs. In this lesson, we will examine variable selection techniques, and how to apply them to select the most informative features for our models.

We will also look at how to effectively save a trained machine

learning model and load it for future use, ensuring it can be embedded into real-world systems, shared, or deployed without the need to retrain.

7 **Regularisation**

egularisation is an important element in data science, used to create models that generalise well and can predict accurately on unseen data. It's a critical defence against overfitting, ensuring models remain adaptable and reliable.

In this lesson, we'll unravel the mechanics and benefits of regularisation methods, including ridge and LASSO regression. We'll delve into the intricacies of data scaling, overfitting, and the strategic application of regularisation to enhance model performance in real-world data science scenarios.

8 **Decision trees**

Understanding decision trees and their implementation is crucial for anyone interested in machine learning and data analysis. Decision trees provide a simple, yet powerful, tool for both classification and regression tasks, allowing us to interpret and visualise complex decision-making processes.

In this lesson, we will delve into the fundamentals of decision trees, exploring how they work, how to train them effectively, and how to implement them using Python libraries like sklearn. Through a combination of theoretical explanations, practical examples, and hands-on coding exercises, we will gain a comprehensive understanding of decision trees and their application in real-world scenarios.

9 **Ensemble methods and bootstrapping**

In the ever-evolving field of data science, ensemble methods stand out as a powerful strategy to improve the predictive performance of machine learning models. These techniques involve combining multiple models to form a stronger, more accurate prediction model. The rationale behind this approach is grounded in the wisdom of crowds theory, which suggests that the aggregate of multiple predictions will often be more accurate than individual predictions.

This lesson delves into ensemble methods and bootstrapping, crucial techniques for anyone looking to enhance their machine learning skills. By mastering these methods, data scientists can significantly improve the robustness, accuracy, and generalisability of their predictive models, addressing some of the most challenging problems in data science today.

10 **Decision tree code challenge**

In this code challenge, we will test our knowledge of the fundamental concepts of decision trees by implementing a decision tree regression model and analysing its RMSLE.