Data Wrangling and Visualization for Data Analyst

No. of Days: 03 (24 hours)

Module 1: Introduction to Data Wrangling and Visualization

- Understanding the data science workflow
    - Overview of the data science process
    - Key steps in the data science workflow
    - Importance of data wrangling and visualization in the workflow

- Overview of data wrangling and visualization concepts
    - Definition and purpose of data wrangling
    - Importance of data cleaning, transformation, and integration
    - Introduction to data visualization and its benefits
    - Exploratory data analysis and storytelling through visualization

- Introduction to popular tools: R, Python, SQL, Power BI, Tableau, and Java
    - Python: Introduction to Python programming language and its data manipulation and visualization libraries
    - SQL: Overview of SQL for data querying and manipulation
    - Power BI: Introduction to Microsoft Power BI for interactive data visualization and reporting
    - Tableau: Overview of Tableau for creating interactive and dynamic visualizations
    - Java: Introduction to Java programming language and its data processing capabilities
- **Lab/Exercise:** Setting up the development environment and exploring the tools
    - In this Lab/Exercise section, you will focus on setting up your development environment and exploring various tools for data wrangling and visualization. The section will cover Python, SQL, Power BI, Tableau, and Java, providing an introduction to each tool and its relevance in data manipulation and visualization.

Module 2: Data Acquisition and Extraction

Objective: In Module 2, you will learn about data acquisition and extraction techniques. You will explore various data sources such as databases, APIs, and web scraping. You will understand how to gather structured data from databases, retrieve data through APIs, and use web scraping to extract information. The module also covers using SQL queries for data extraction and introduces data extraction using programming languages like R and Python. Finally, you will have a lab/exercise session to practice extracting data from different sources using SQL, R, and Python.

- Gathering data from various sources: databases, APIs, web scraping
  - Data sources overview: Databases, APIs, web scraping
  - Extracting structured data from databases
  - Retrieving data from APIs
  - Web scraping techniques for data extraction

- Extracting data using SQL queries
  - Introduction to SQL for data extraction
  - Writing SQL queries to extract data from databases
  - Filtering, joining, and aggregating data with SQL
- Extracting data using programming languages: R and Python
  - Introduction to data extraction in R
  - Extracting data from databases using R
  - Introduction to data extraction in Python
  - Extracting data from databases using Python

- **Lab/Exercise:** Extracting data from different sources using SQL, R, and Python
  - Hands-on exercise to practice data extraction from databases using SQL
  - Lab session on extracting data using R programming language
  - Lab session on extracting data using Python programming language

Module 3: Data Cleaning and Preprocessing

Objective: In the Lab/Exercise section of Module 3, you will focus on hands-on activities related to data cleaning and preprocessing. You will set up your development environment and explore the use of R, Python, and SQL for cleaning and preprocessing data. The exercises will involve applying techniques to handle missing values, outliers, and inconsistent data. Additionally, you will practice data transformation techniques such as normalization, scaling, encoding categorical variables, and feature engineering using the respective programming languages.

- Exploring and understanding the data
  - Overview of data exploration and analysis
  - Techniques for understanding data distributions, patterns, and relationships

- Handling missing values, outliers, and inconsistent data
  - Identifying missing values and their impact on analysis
  - Strategies for handling missing values: imputation, deletion, or modelling
  - Detecting and managing outliers in the data
  - Dealing with inconsistent or erroneous data entries

- **Lab/Exercise:** Cleaning and preprocessing data using R, Python, and SQL
  - Setting up the development environment

- o Cleaning and preprocessing data using R
- o Cleaning and preprocessing data using Python
- o Cleaning and preprocessing data using SQL

Module 4: Data Transformation and Feature Engineering

Objective: In the Lab/Exercise section of Module 4, you will focus on hands-on activities related to data transformation and feature engineering. You will set up your development environment and explore the use of R, Python, and SQL for performing data transformation and feature engineering tasks. The exercises will involve applying feature selection and extraction techniques, creating derived variables, aggregating data, and applying transformations and normalization methods to improve data quality.

- Feature selection and extraction techniques
  - o Importance of feature selection and extraction in machine learning
  - o Common feature selection methods: filtering, wrapper, and embedded approaches
  - o Feature extraction techniques: PCA and statistical measure.

- Creating derived variables and aggregating data
  - o Generating new variables from existing ones: mathematical operations, transformations, and interactions
  - o Aggregating data at different levels: group by, pivot tables, and summary statistics

- Applying transformations and normalization to improve data quality
  - o Data transformation techniques: logarithmic, exponential, and power transformations
  - o Normalization methods: min-max scaling, z-score normalization, and more
  - o Handling skewed data and outliers through transformations and normalization

- **Lab/Exercise:** Performing data transformation and feature engineering using R, Python, and SQL
  - o Setting up the development environment
  - o Data transformation using R
  - o Data transformation using Python
  - o Data transformation using SQL

Module 5: Exploratory Data Analysis (EDA)

Objective: In the Lab/Exercise section of Module 5, you will focus on hands-on activities related to exploratory data analysis (EDA) and creating visualizations. You will set up your development environment and explore the use of R, Python, Power BI, and Tableau for conducting EDA and creating visualizations. The exercises will involve applying data visualization principles and best practices to create informative charts, graphs, and plots. Additionally, you will explore relationships and patterns in the data through visual analysis.

- Data visualization principles and best practices
  - Importance of data visualization in EDA
  - Principles of effective data visualization
  - Best practices for creating clear and informative visualizations

- Creating informative visualizations using charts, graphs, and plots
  - Overview of commonly used visualization types: bar charts, line charts, scatter plots, histograms, etc.
  - Choosing the appropriate visualization for different data types and objectives
  - Enhancing visualizations with labels, colours, and annotations

- Exploring relationships and patterns in the data
  - Identifying correlations and dependencies
  - Uncovering trends and patterns through visual analysis
  - Analyzing distributions and outliers in the data

- **Lab/Exercise:** Conducting EDA and creating visualizations using R, Python, Power BI, and Tableau
  - Setting up the development environment
  - Exploratory Data Analysis using R
  - Exploratory Data Analysis using Python
  - Creating visualizations using Power BI
  - Creating visualizations using Tableau

Module 6: Advanced Data Visualization Techniques

<u>Objective</u>: In the Lab/Exercise section of Module 6, you will focus on hands-on activities related to advanced data visualization techniques. You will set up your development environment and explore the use of Power BI, Tableau, and Python libraries for building interactive visualizations and dashboards. The exercises will involve creating interactive visualizations, exploring geospatial data, visualizing network structures, and performing social network analysis.

- Interactive visualizations and dashboards
  - Importance of interactive visualizations in data analysis
  - Creating interactive visualizations using interactivity features
  - Designing and building interactive dashboards for data exploration

- Geospatial data visualization
  - Overview of geospatial data and its visualization
  - Mapping techniques and tools for visualizing geographic data
  - Adding layers, markers, and tooltips to geospatial visualizations

- Network visualization and social network analysis
  - Understanding network data and its visualization
  - Techniques for visualizing network structures and relationships

   o Introduction to social network analysis and its visualization

- **Lab/Exercise:** Building interactive visualizations and dashboards using Power BI, Tableau, and Python libraries
  - Setting up the development environment
  - Building interactive visualizations and dashboards using Power BI
  - Building interactive visualizations and dashboards using Tableau
  - Building interactive visualizations using Python libraries

Module 7: Presenting Insights and Reporting

Objective: In the Lab/Exercise section of Module 7, you will focus on hands-on activities related to presenting insights and reporting. You will explore the use of Power BI, Tableau, and presentation tools to design and present data-driven insights. The exercises will involve creating compelling data presentations and reports, incorporating storytelling techniques, and effectively communicating insights and recommendations to different stakeholders.

- Creating compelling data presentations and reports
  - Importance of effective data presentations and reports
  - Principles of creating visually appealing and impactful presentations
  - Designing clear and concise data-driven reports

- Storytelling with data
  - Using storytelling techniques to convey insights
  - Structuring a narrative for data-driven storytelling
  - Incorporating visuals and narratives to enhance storytelling
- Effective communication of insights and recommendations
  - Communicating complex ideas and findings in a simple manner
  - Tailoring messages to different stakeholders
  - Presenting actionable insights and recommendations

- **Lab/Exercise:** Designing and presenting data-driven insights using Power BI, Tableau, and presentation tools
  - Setting up the development environment
  - Designing data presentations using Power BI
  - Designing data presentations using Tableau
  - Presenting insights using presentation tools

==