Postgraduate Certificate in Data-Driven Science Journalism

Data Wrangling

Data wrangling is an essential process in data science and journalism that involves cleaning, transforming, and organizing raw data into a format suitable for analysis. In this course, we will explore key terms and vocabulary related to data wrangling to help you understand and apply these concepts effectively in your work as a data-driven science journalist.

Data Wrangling: Data wrangling, also known as data munging, is the process of cleaning and transforming raw data into a more structured format for analysis. It involves tasks such as removing duplicates, handling missing values, and converting data types.

Data Cleaning: Data cleaning is the process of detecting and correcting errors in a dataset. This may involve removing duplicates, handling missing values, correcting inconsistencies, and standardizing data formats.

Data Transformation: Data transformation involves converting data from one format to another. This may include aggregating data, creating new variables, or reshaping data to make it more suitable for analysis.

Data Organization: Data organization involves structuring data in a way that makes it easier to analyze. This may involve sorting data, grouping data into categories, or creating hierarchies.

Data Quality: Data quality refers to the accuracy, completeness, and reliability of data. High-quality data is essential for producing accurate and reliable analyses.

Data Preprocessing: Data preprocessing involves preparing raw data for analysis. This may include data cleaning, data transformation, and data organization.

Missing Values: Missing values are data points that are not recorded or available in a dataset. Handling missing values is an important part of data wrangling, as they can affect the accuracy of analyses.

Data Types: Data types refer to the format in which data is stored, such as integers, strings, or dates. Understanding data types is important for data wrangling, as different data types require different handling.

Outliers: Outliers are data points that are significantly different from the rest of the dataset. Identifying and handling outliers is important in data wrangling to prevent them from skewing analyses.

Aggregation: Aggregation involves combining and summarizing data to create new insights. This may involve calculating averages, sums, or counts of data points.

Normalization: Normalization is the process of scaling data so that all values fall within a specific range. Normalizing data is important for ensuring that different variables are comparable.

Reshaping: Reshaping data involves changing the structure of a dataset. This may include pivoting data, melting data, or transposing data to make it more suitable for analysis.

Joining: Joining involves combining data from multiple datasets based on a common key. There are different types of joins, such as inner joins, outer joins, left joins, and right joins.

Filtering: Filtering involves selecting a subset of data based on specific criteria. This may involve removing rows or columns that do not meet certain conditions.

Grouping: Grouping involves aggregating data based on specific variables. This may involve calculating summary statistics for each group or creating groups for further analysis.

Regular Expressions: Regular expressions are sequences of characters that define a search pattern. They are commonly used in data wrangling to extract, replace, or manipulate text data.

APIs: APIs, or Application Programming Interfaces, allow different software applications to communicate with each other. APIs are commonly used in data wrangling to access and retrieve data from external sources.

Web Scraping: Web scraping is the process of extracting data from websites. Web scraping is a useful technique in data wrangling for collecting data from online sources.

Challenges in Data Wrangling: Data wrangling can be a complex and time-consuming process. Some common challenges include dealing with missing values, handling outliers, merging datasets, and ensuring data quality.

Practical Applications of Data Wrangling: Data wrangling is essential for any data-driven science journalist. It allows you to clean and prepare data for analysis, identify trends and patterns, and communicate insights effectively to your audience.

By mastering the key terms and vocabulary related to data wrangling, you will be better equipped to tackle real-world data challenges and produce high-quality data-driven journalism.

Data wrangling is a crucial process in the field of data science and journalism that involves cleaning, transforming, and organizing raw data into a format that is suitable for analysis. This process is essential for ensuring the accuracy and reliability of the data before it can be used to draw meaningful insights and conclusions. In this course, you will learn about key terms and vocabulary related to data wrangling that will help you navigate and manipulate datasets effectively.

Data Wrangling: Data wrangling, also known as data munging, is the process of cleaning, structuring, and enriching raw data into a format suitable for analysis.

Data Cleaning: Data cleaning involves identifying and correcting errors or inconsistencies in the data to improve its quality and accuracy. This may include handling missing values, removing duplicates, and correcting formatting issues.

Data Transformation: Data transformation involves converting data from one format to another, such as changing data types, aggregating data, or creating new variables based on existing ones.

Data Organization: Data organization involves structuring the data in a way that makes it easy to work with, such as sorting data, grouping data into categories, and creating hierarchies.

Data Quality: Data quality refers to the accuracy, completeness, consistency, and reliability of the data. High-quality data is essential for making informed decisions and drawing accurate conclusions.

Data Integration: Data integration involves combining data from multiple sources into a single dataset for analysis. This may require aligning data formats, resolving inconsistencies, and merging datasets.

Data Validation: Data validation is the process of ensuring that the data meets specific criteria or standards. This may involve checking for errors, outliers, or inconsistencies in the data.

Data Exploration: Data exploration involves analyzing the data to understand its characteristics, patterns, and relationships. This may include visualizing the data, calculating summary statistics, and identifying trends.

Data Visualization: Data visualization is the graphical representation of data to communicate insights effectively. This may include charts, graphs, maps, and other visualizations to help interpret and present data.

Data Analysis: Data analysis involves using statistical methods and algorithms to extract meaningful insights from the data. This may include hypothesis testing, regression analysis, clustering, or machine learning techniques.

Data Manipulation: Data manipulation involves changing the structure or content of the data to perform specific tasks, such as filtering, sorting, merging, or reshaping the data.

Data Extraction: Data extraction involves retrieving data from various sources, such as databases, APIs, websites, or files, to create a dataset for analysis.

Data Aggregation: Data aggregation involves combining individual data points into summary statistics, such as averages, totals, or counts, to analyze trends and patterns in the data.

Data Mining: Data mining is the process of discovering patterns, trends, and insights in large datasets using machine learning algorithms, statistical techniques, or other data analysis methods.

Data Governance: Data governance refers to the management and control of data assets within an organization to ensure data quality, security, and compliance with regulations.

Data Privacy: Data privacy involves protecting the confidentiality and security of personal or sensitive data to prevent unauthorized access or misuse.

Data Ethics: Data ethics refers to the moral and ethical considerations surrounding the collection, use, and dissemination of data, including issues of bias, privacy, transparency, and accountability.

Data Security: Data security involves protecting data from unauthorized access, disclosure, or modification to ensure the confidentiality, integrity, and availability of the data.

Data Storage: Data storage refers to the physical or digital locations where data is stored, such as databases, cloud storage, servers, or external drives.

Data Structure: Data structure refers to the organization and format of the data, including how data is stored, accessed, and manipulated. Common data structures include tables, lists, arrays, and trees.

Data Model: A data model is a representation of the structure, relationships, and constraints of the data, such as entity-relationship models, relational models, or object-oriented models.

Data Schema: A data schema is a blueprint or description of the structure and properties of a dataset, including data types, field names, and relationships between data elements.

ETL (Extract, Transform, Load): ETL is a process used to extract data from various sources, transform it into a consistent format, and load it into a target database for analysis.

Data Pipeline: A data pipeline is a series of interconnected steps or processes that move data from its source to its destination, including data extraction, transformation, and loading.

Data Source: A data source is the origin or location of the data, such as a database, file, API, sensor, or website, where data can be retrieved for analysis.

Data Format: Data format refers to the structure and encoding of the data, such as CSV, JSON, XML, or Parquet, which determines how the data is stored and processed.

Data Query: A data query is a request or command to retrieve specific data from a database or dataset, using SQL, NoSQL, or other query languages to filter, sort, or aggregate data.

Data Migration: Data migration is the process of transferring data from one system to another, such as upgrading software, moving to a new platform, or consolidating data sources.

Data Cleansing: Data cleansing is the process of identifying and correcting errors, inconsistencies, or missing values in the data to improve its quality and reliability.

Data Enrichment: Data enrichment is the process of enhancing the data with additional information, such as geolocation data, demographic data, or social media data, to create more valuable insights.

Data Anomalies: Data anomalies are irregularities or deviations in the data that do not conform to the expected patterns, such as errors, outliers, duplicates, or missing values.

Data Profiling: Data profiling is the process of analyzing the structure and content of the data to understand its characteristics, quality, and patterns before performing further analysis.

Data Governance Framework: A data governance framework is a set of policies, processes, and procedures that define how data is managed, controlled, and protected within an organization.

Data Dictionary: A data dictionary is a document or database that describes the structure, elements, and definitions of the data, including metadata, data types, and relationships between data elements.

Data Catalog: A data catalog is a centralized repository or database that stores metadata and information about available datasets, including data sources, descriptions, and usage guidelines.

Data Lake: A data lake is a centralized repository that stores a vast amount of raw data in its native format, allowing for flexible analysis and exploration of the data using various tools and technologies.

Data Warehouse: A data warehouse is a centralized repository that stores structured, integrated, and historical data for reporting, analysis, and decision-making purposes in an organization.

Data Silos: Data silos are isolated or disconnected storage systems or databases that hinder the sharing, integration, or access of data across different departments or systems within an organization.

Data Architecture: Data architecture is the design and structure of the data systems, including databases, storage, processing, and integration components, to support data management and analysis.

Data Strategy: A data strategy is a plan or roadmap that outlines how data will be collected, stored, analyzed, and used to achieve business goals or objectives within an organization.

Big Data: Big data refers to large and complex datasets that are difficult to process and analyze using traditional data management tools and techniques, requiring specialized technologies and methods.

Data Science: Data science is an interdisciplinary field that combines statistics, mathematics, computer science, and domain knowledge to extract insights, patterns, and knowledge from data using various techniques and algorithms.

Machine Learning: Machine learning is a subset of artificial intelligence that involves developing algorithms and models that can learn from data, identify patterns, and make predictions or decisions without explicit programming.

Structured Data: Structured data is organized and formatted data that is stored in a tabular format, such as databases, spreadsheets, or CSV files, with a clear schema and well-defined relationships between data elements.

Unstructured Data: Unstructured data is data that does not have a predefined format or structure, such as text, images, videos, or social media posts, making it difficult to analyze using traditional methods.

Semi-Structured Data: Semi-structured data is data that has a flexible schema or partial organization, such as JSON, XML, or NoSQL databases, allowing for more flexibility in storing and manipulating data.

Data Wrangling Tools: Data wrangling tools are software or platforms that help automate and streamline the data cleaning, transformation, and organization processes, such as Excel, Python, R, SQL, or commercial data preparation tools.

Data Wrangling Challenges: Data wrangling challenges include dealing with missing data, handling outliers, resolving inconsistencies, merging datasets, detecting errors, and ensuring data quality and integrity throughout the process.

Data Wrangling Best Practices: Data wrangling best practices include documenting data transformations, using version control, automating repetitive tasks, validating data quality, and collaborating with stakeholders to ensure data accuracy and reliability.

Real-World Applications: Data wrangling is used in various industries and domains, such as finance, healthcare, marketing, e-commerce, social media, and government, to clean, transform, and analyze data for decision-making, reporting, and research purposes.

Conclusion: Data wrangling is a fundamental process in data-driven science journalism that involves cleaning, transforming, and organizing raw data into a format suitable for analysis. By understanding key terms and vocabulary related to data wrangling, you will be able to effectively navigate and manipulate datasets to uncover meaningful insights and trends in the data.