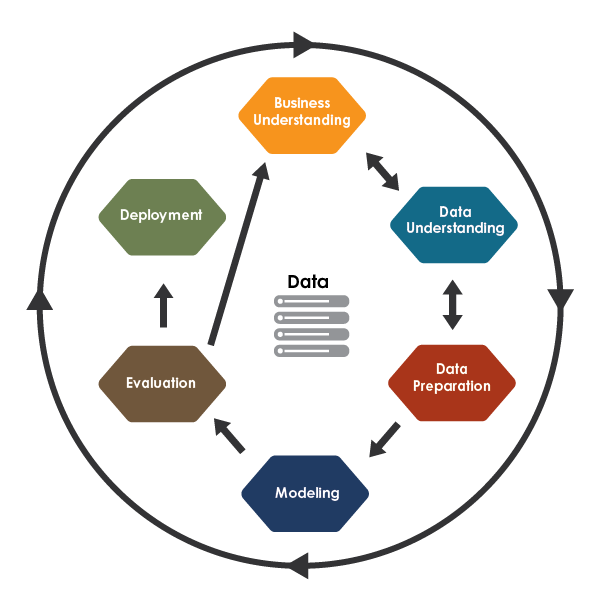
**CRISP-DM**



**Introduction**

In today’s data-driven world, we must navigate a vast sea of information to extract valuable insights. In order to guide us safely through these challenging waters, we require a reliable compass: the data science workflow.

**What is the data science workflow?**

The data science workflow is a structured framework of stages that guides data scientists in effectively navigating the complexities of data science projects.

**Stages**

**1) Definition  
2) Collection  
3) Preparation  
4) Exploration  
5) Analysis  
6) Communication**

**Importance**

The data science workflow empowers data scientists to collaborate efficiently and effectively when extracting value from data.

**Challenges**

The data science workflow is inherently iterative, so it is crucial to recognise the need to revisit earlier stages when new insights emerge.

**Alternative Frameworks**

There is no one-size-fits-all data science workflow, accordingly this article offers a personalised take, drawing inspiration from widely recognised frameworks like [CRISP-DM](https://www.datascience-pm.com/crisp-dm-2/) and [OSEMN](https://machinelearningmastery.com/how-to-work-through-a-problem-like-a-data-scientist/).

**1) Definition**

The definition stageinvolves clearly outlining the project in order to ensure that efforts, expectations, and resources are aligned with a shared purpose and direction.

**Techniques**

**Context**Gather contextual information related to the project (e.g. causes, goals, issues, expectations, implications)

**Objectives**Define desired outcomes, measurable goals, and key questions before breaking tasks into distinct, manageable components

**Constraints**Determine the limitations of the project by considering important factors (e.g. resource availability, time constraints, data accessibility, ethical considerations)

**2) Collection**

The collection stage involves acquiring the necessary data in order to perform a meaningful analysis based upon accurate information.

**Techniques**

**Data Requirements**Define which data is needed to properly approach the project (e.g. format, variables, time range, granularity)

**Data Sources**Find reliable and relevant data sources (e.g. databases, APIs, files, sensor readings)

**Authentication**Secure necessary permissions to access the data (e.g. email/password, OAuth, API key, robots.txt)

**Collection**Acquire the data using appropriate methods (e.g. SQL queries, API calls, web scraping, manual data entry)

**Data Management**Handle the data in accordance with best practices (e.g. data quality, data governance, data security)

**3) Preparation**

The preparation stage involves processing the raw data in order to achieve a consistent and structured format that is well-suited for a reliable analysis.

**Techniques**

**Data Cleaning**Identify and handle errors and inconsistencies in the data (e.g. missing values, duplicate entries, anomalies, data formats)

**Data Integration**Combine data from multiple sources whilst ensuring consistency (e.g. variables, naming conventions, indexing)

**Feature Engineering**Engineer meaningful features from raw data (e.g. feature selection, feature creation, data transformation)

**4) Exploration**

The exploration stage involves understanding the main characteristics of the data in order to formulate valid hypotheses, identify issues, and refine the project definition.

**Techniques**

**Distribution Analysis**Examine the distribution of each variable (e.g. mean, median, standard deviation, skew, outliers)

**Dependency Analysis**Investigate and quantify variable relationships to understand how they influence each other (e.g. correlations, interactions, covariances, time series analysis)

**Data Segmentation**Explore the data using various segments and subsets to understand how patterns vary across different groups

**Hypothesis Generation**Generate initial insights to develop hypotheses about relationships and patterns

**5) Analysis**

The analysis stage involves performing an in-depth examination of the data in order to develop a robust solution that is capable of producing valuable insights.

**Techniques**

**Hypothesis Testing**Apply significance tests to assess the statistical importance of observed patterns and relationships (e.g. t-test, ANOVA, chi-squared test)

**Advanced Techniques**Utilise advanced algorithms relevant to specific hypotheses (e.g. time series analysis, regression analysis, anomaly detection)

**Modelling**Select, build, and assess suitable models with relevant metrics to identify the optimal configuration whilst considering trade-offs such as complexity, interpretability, and performance

**6) Communication**

The communication stage involves presentingthe project and its findings to stakeholders in order to create clarity and awareness.

**Techniques**

**Model Deployment**Deploy the model for real-world use (e.g. create an API, build a web application, integrate into an existing system)

**Monitoring and Logging**Implement performance tracking and issue logging for the model during usage

**Documentation**Create comprehensive project documentation covering technical details (e.g. model architecture, data sources, assumptions, limitations)

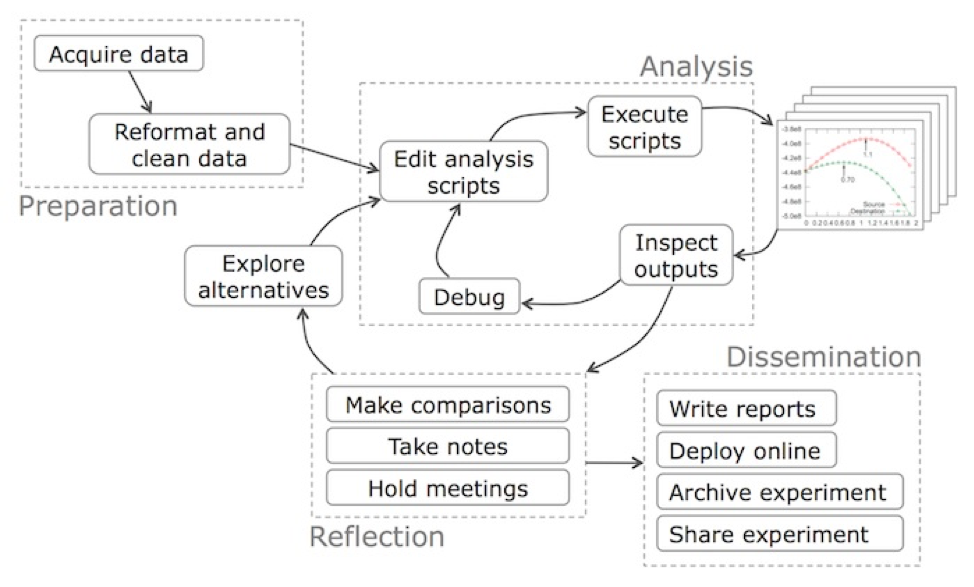
**Reporting and Presentation**Produce and deliver concise, informative, and engaging project summaries (e.g. objectives, methods, results, insights, key findings)

**Conclusion**

The data science workflow is an essential tool because it provides structure and organisation to complex projects, resulting in improved decision-making, enhanced collaboration, and greater accuracy.

Data science is a dynamic field, and whilst the workflow provides a solid foundation, it should be adapted to fit specific project needs and goals.

Embracing and applying the data science workflow will empower data scientists to streamline their process and thrive in the ever-changing, ever-growing sea of data.

+

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Harvard** | **CRISP-DM** | **OSEMN** | **Guo’s** | **Tandel’s** | **Joshi’s** |
| **Understand** | Ask an interesting question | Business understanding |  |  | Understand the objective |  |
| **Acquire** | Get the data | Data understanding | Obtain | Prepare (acquire) | Import the data | Find, connect and access data |
| **Clean** |  | Data preparation | Scrub | Prepare (clean) | Clean the data | Prepare the data |
| **Explore** | Explore the data |  | Explore |  | Explore the data |  |
| **Model** | Model the data | Modeling | Model | Analysis | Model the data | Build models |
| **Evaluate** |  | Evaluation | iNterpret | Reflect |  |  |
| **Communicate** | Communicate / Visualize |  |  | Disseminate | Communicate results |  |
| **Deploy** |  | Deployment |  |  |  | Deploy models |
| **Monitor** |  |  |  |  |  | Monitor models |