Data science process:

Here's the end-to-end data science process! Following a structured approach to data science helps you to maximize your chances of success in a data science project at the lowest cost. It also makes it possible to take up a project as a team, with each team member focusing on what they do best. The figure summarizes the data science process and shows the main steps and actions you'll take during a project.

Data science process

- **1: Setting the research goal**
  - Define research goal
  - Create project charter

- **2: Retrieving data**
  - Internal data
    - Data retrieval
    - Data ownership
  - External data

- **3: Data preparation**
  - Data cleansing
    - Errors from data entry
    - Physically impossible values
    - Missing values
    - Outliers
    - Spaces, typos, …
    - Errors against codebook
  - Data transformation
    - Aggregating data
    - Extrapolating data
    - Derived measures
    - Creating dummies
    - Reducing number of variables
  - Combining data
    - Merging/joining data sets
    - Set operators
    - Creating views

- **4: Data exploration**
  - Simple graphs
  - Combined graphs
  - Link and brush
  - Nongraphical techniques

- **5: Data modeling**
  - Model and variable selection
  - Model execution
  - Model diagnostic and model comparison

- **6: Presentation and automation**
  - Presenting data
  - Automating data analysis

Manning.com, Via: Machine Learning India – aml.india