Reference guide: The EDA process

The six practices of EDA are iterative and non-sequential

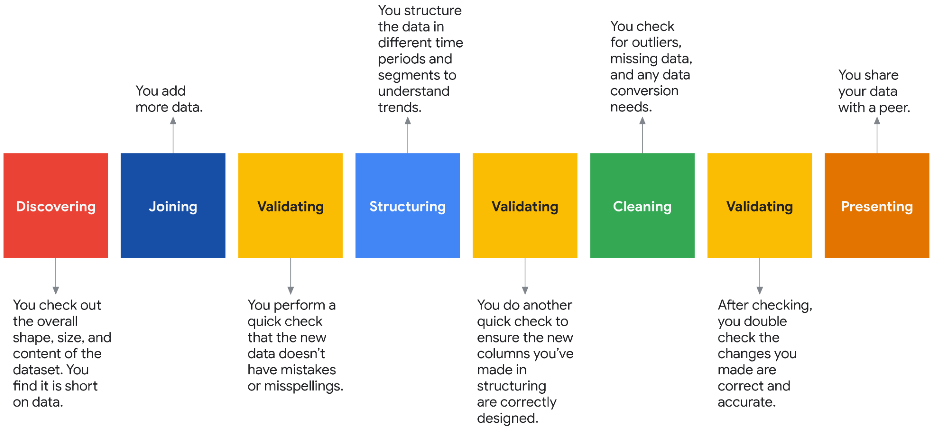
Exploratory data analysis (EDA) is not like a cake recipe. It is ***not*** a step-by-step process you follow. Instead, the six practices of EDA are iterative and non-sequential.

- **Iterative**: Relating to or involving repetition of a process
- **Non-sequential**: Not arranged in or following an order or sequence.

Because of the varying nature of datasets, the approach to exploring that data will be different each time. That means that you will need to use your logic and experience throughout the EDA process to determine which of the six practices to utilize, how many times to apply them, and when in the process you should apply them.

Visual example

Imagine you are assigned a dataset that has only 200 rows and five columns of data about trees in a coniferous forest in Norway. You know that to complete your full analysis you'll need more than 1,000 rows and at least two more columns. Even without much more detail than that, your entire EDA process might look something like this:


You structure the data in different time periods and segments to understand trends.
You check for outliers, missing data, and any data conversion needs.
You add more data.
You share your data with a peer.
Discovering
Joining
Validating
Structuring
Validating
Cleaning
Validating
Presenting
You check out the overall shape, size, and content of the dataset. You find it is short on data.
You perform a quick check that the new data doesn't have mistakes or misspellings.
You do another quick check to ensure the new columns you've made in structuring are correctly designed.
After checking, you double check the changes you made are correct and accurate.

1. **Discovering**: You check out the overall shape, size, and content of the dataset. You find it is short on data.

2. **Joining**: You add more data.
3. **Validating**: You perform a quick check that the new data doesn't have mistakes or misspellings.
4. **Structuring**: You structure the data in different time periods and segments to understand trends.
5. **Validating:** You do another quick check to ensure the new columns you've made in structuring are correctly designed.
6. **Cleaning**: You check for outliers, missing data, and needs for conversions or transformations,
7. **Validating**: After cleaning, you double check the changes you made are correct and accurate,
8. **Presenting**: You share your dataset with a peer.

Notice you performed the "validating" practice iteratively, or multiple times, to make sure your changes to the data did not unwittingly introduce errors. Also, because you recognized the need for more data up front, the practice of "joining" was performed immediately following the practice of "discovering."

After you present your cleaned dataset to a peer, there is a good chance you will receive notes or ideas for more exploration and/or cleaning. Because of that, you will see even more iterations.

Pro tip: Data scientists expect to perform the practices of EDA multiple times on a dataset before they feel comfortable declaring it "clean" and ready for modeling or machine learning algorithms.

The importance of EDA in ethical machine learning

As algorithms and machine learning networks begin to make more and more decisions on behalf of individuals, companies, and even governments, the discussion of ethics and regulation becomes more and more important. According to the Institute for Ethical AI & Machine Learning, there are eight principles for developing machine learning systems in a responsible way.

Key principles of the EDA process

The following two principles are inherently part of the EDA process:

- **Human augmentation**: This principle ensures humans are inserted throughout the AI or machine learning algorithm systems for oversight. Thorough EDA, performed by data scientists, is perhaps one of the best ways to limit bias, imbalance, and inaccuracies being fed into an algorithm.

- **Bias evaluation**: Without human interference, bias is too easily injected and reproduced in machine learning models. Performing methodical EDA processes will lead data scientists to be aware of and act on biases and imbalances in the data.

Pro tip: The importance of assuring adherence to ethical standards cannot be overstated in the data career space. Data professionals need to continuously grow their capacities to recognize bias and discrimination by consistently applying an ethical mindset to their EDA work.

Beyond machine learning, EDA is applicable to nearly any important data-based decision. Moving forward, you will learn about many applications of EDA and the necessity of an iterative and non-sequential approach.