**Three Critical Elements of Data Preprocessing — Part 1**

The backbone of modeling in data science.

In this article, I will share the main components of the data preprocessing step in a data science project life cycle and provide some valuable resources to perform this step efficiently.

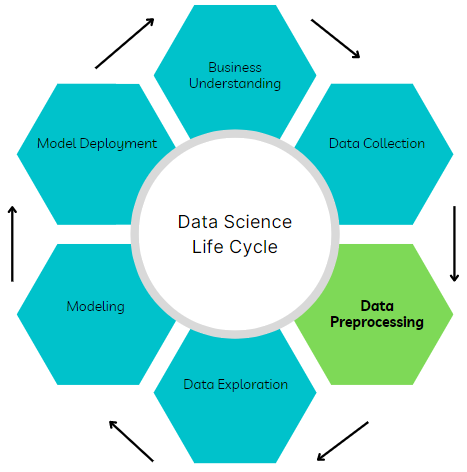


Image by author

Data preprocessing is a critical step that converts raw data from different sources into a refined form that can be used to derive actionable insights. It entails integration, cleaning, and transformation.

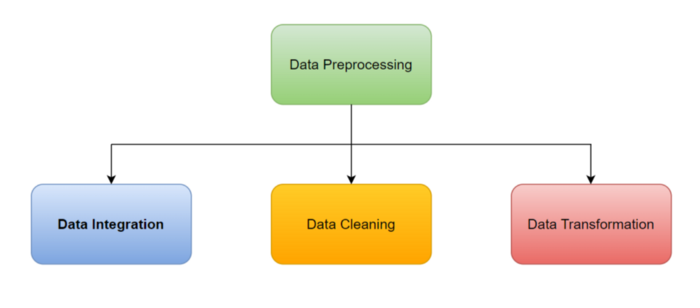


Image by author

Additionally, data preprocessing ensures that quality data is available to machine learning models resulting in excellent prediction performance. In fact, the saying “Garbage in Garbage out” in modeling rests heavily on the quality of data supplied to the models.

Hence, data preprocessing can be described as the backbone of modeling.

The tools of choice for data preprocessing are Pandas and Numpy however, equivalent libraries in other languages may be used and data extraction from databases is mostly done with SQL queries.

Each part of this series will cover one critical element of data preprocessing in detail starting with data integration.

**Data Integration**

This component of data preprocessing involves exploiting the relationships between disparate datasets by combining them using similar features as connection points.

Data integration skills can help data scientists harness informative data available in silos thereby creating more business value from existing resources.

Data integration can be performed by using SQL queries to connect directly to different sources and returning a single dataset consisting of attributes from the data sources.

Alternatively, individual queries may be written to pull data from different sources, and python libraries such as Pandas and Numpy may be used to combine the data to produce the required dataset.

Option 2 from above is my preferred approach based on my work experience and knowledge of Python programming.

The key actions performed during data integration include:

**Join**

This action results in an increase in the number of columns with or without a change in the number of rows of the main data depending on the type of join. For example, this action may be used to enrich the input data to a machine-learning model by adding new features (columns) to existing training data.

In Python, the Pandas’ Concat, Merge and Numpy’s Concatenate modules can be used to perform join operations.

The image below shows the different types of joins:

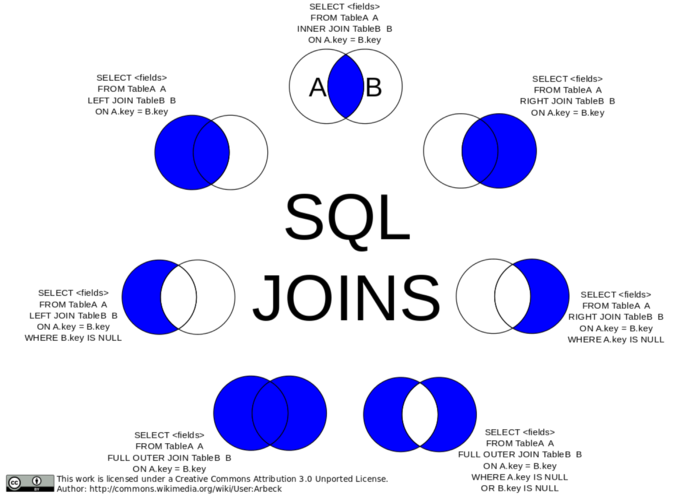


Photo by Arbeck, CC BY 3.0, via Wikimedia

*Resources:*

SQL Join: <https://www.w3schools.com/sql/sql_join.asp>

Pandas Concat: <https://pandas.pydata.org/docs/reference/api/pandas.concat.html>

Pandas Merge: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html>

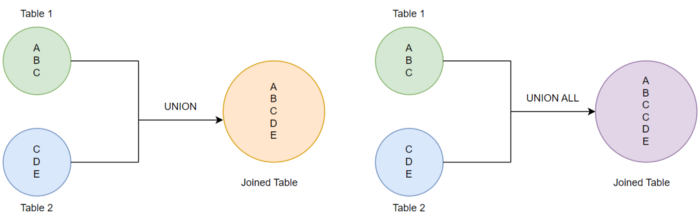
Numpy Concatenate: <https://numpy.org/doc/stable/reference/generated/numpy.concatenate.html>

**Union**

This action increases the number of rows of the main dataset with no change in the number of columns. For example, this action may be used to increase the number of examples available for training a machine learning model to reduce overfitting. All datasets to be unionized must have the same features for the returned dataset to be usable.

Caution must be taken to avoid duplicate data when performing union actions. If permitted, UNION ALL action allows all data from both datasets to be represented in the returned data irrespective of duplicate examples.

Pandas Concat can also be used to perform union operations.



Union vs Union All (Image by author)

*Resources:*

SQL Union:

1. <https://www.w3schools.com/sql/sql_union.asp>
2. <https://www.tutorialspoint.com/sql/sql-unions-clause.htm>

**Conclusions**

In this article, we have explored data integration, a critical component of the data preprocessing step in a data science project life cycle. It involves combining data from different sources to obtain a dataset with all available relevant features and examples.

In the next article, we will cover data cleaning, another critical component of data preprocessing.

I hope you enjoyed reading this article, until next time. Cheers!