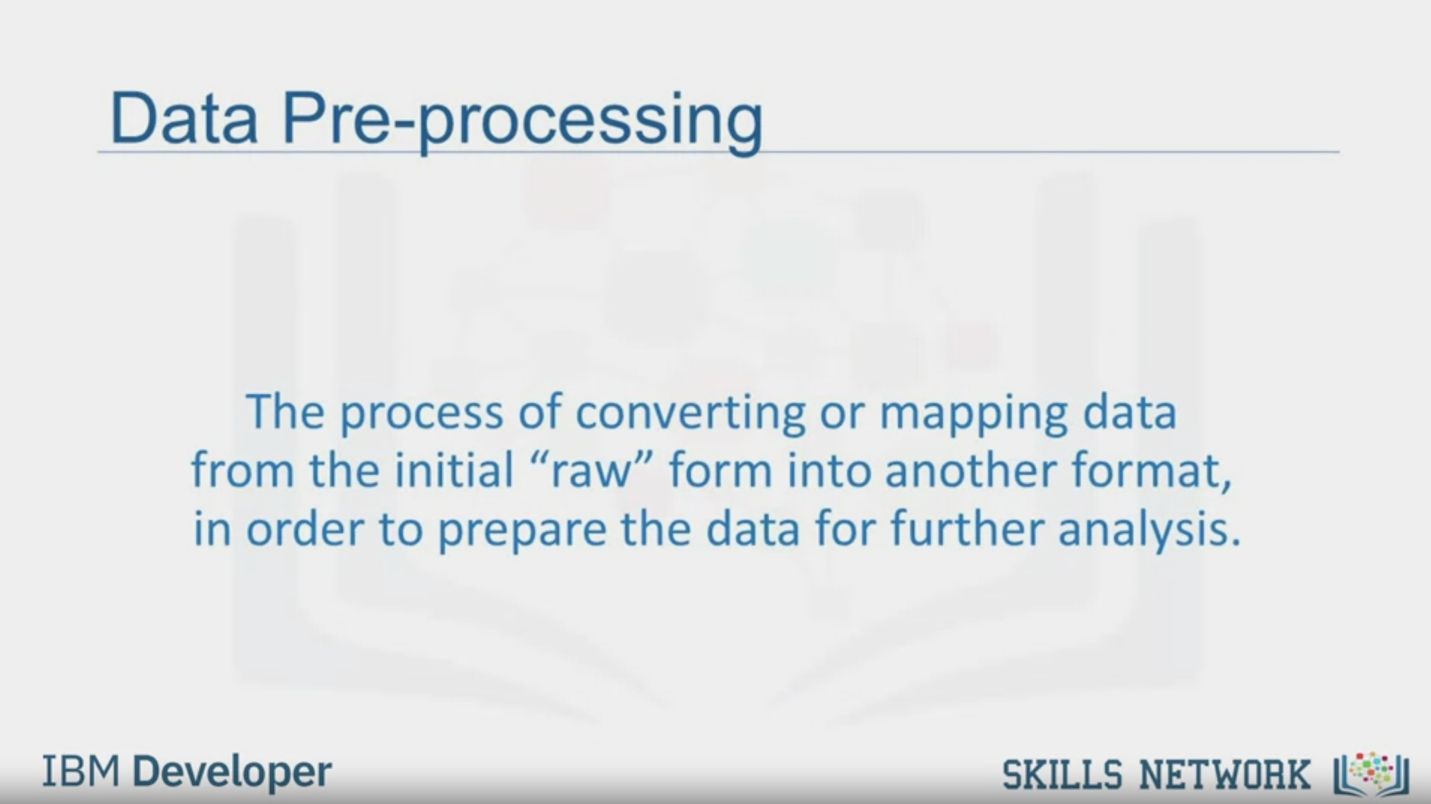


Pre-processing Data in Python

IBM Developer

SKILLS NETWORK



Data Pre-processing

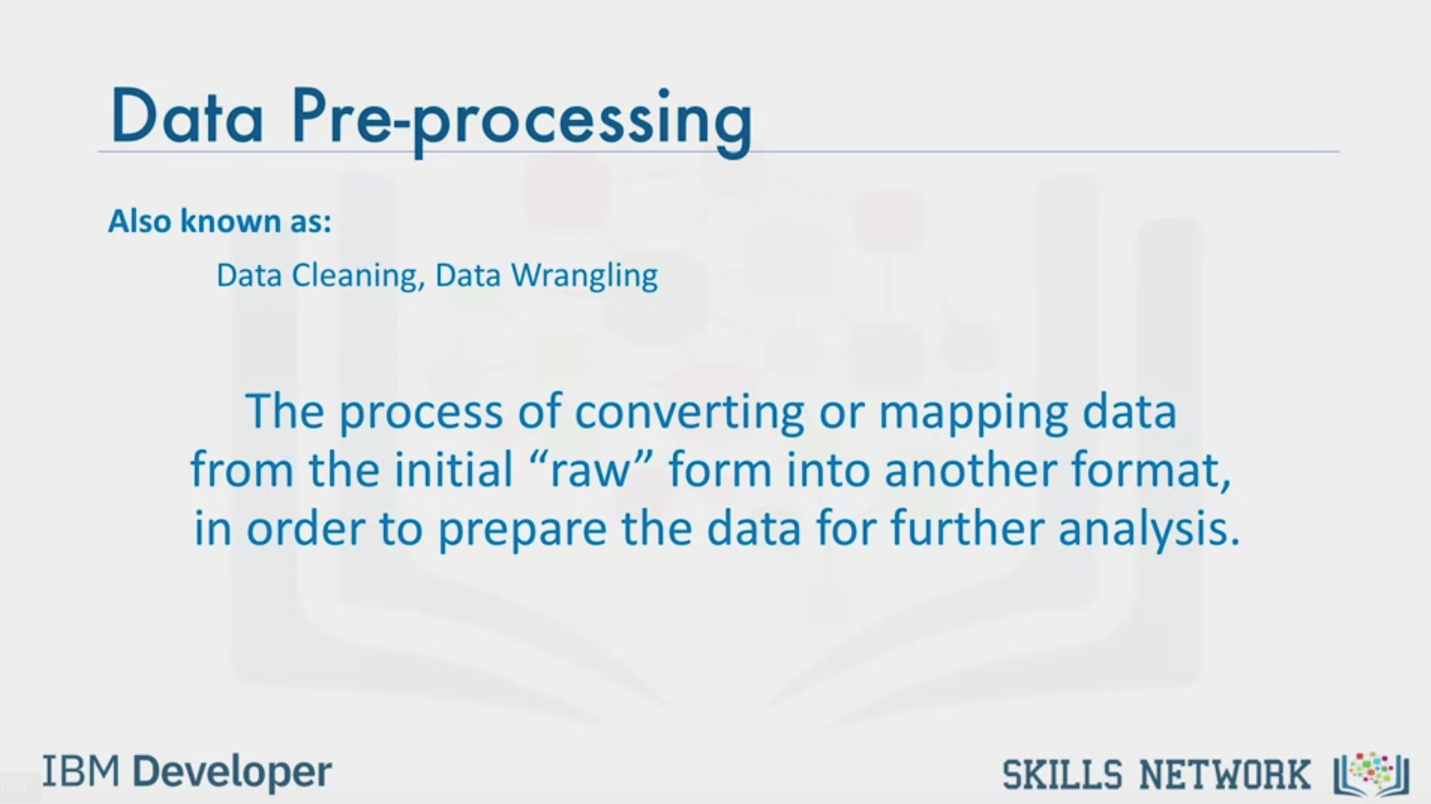
The process of converting or mapping data

from the initial "raw" form into another format.

in order to prepare the data for further analysis.

IBM Developer

SKILLS NETWORK



Data Pre-processing

Also known as:

Data Cleaning, Data Wrangling

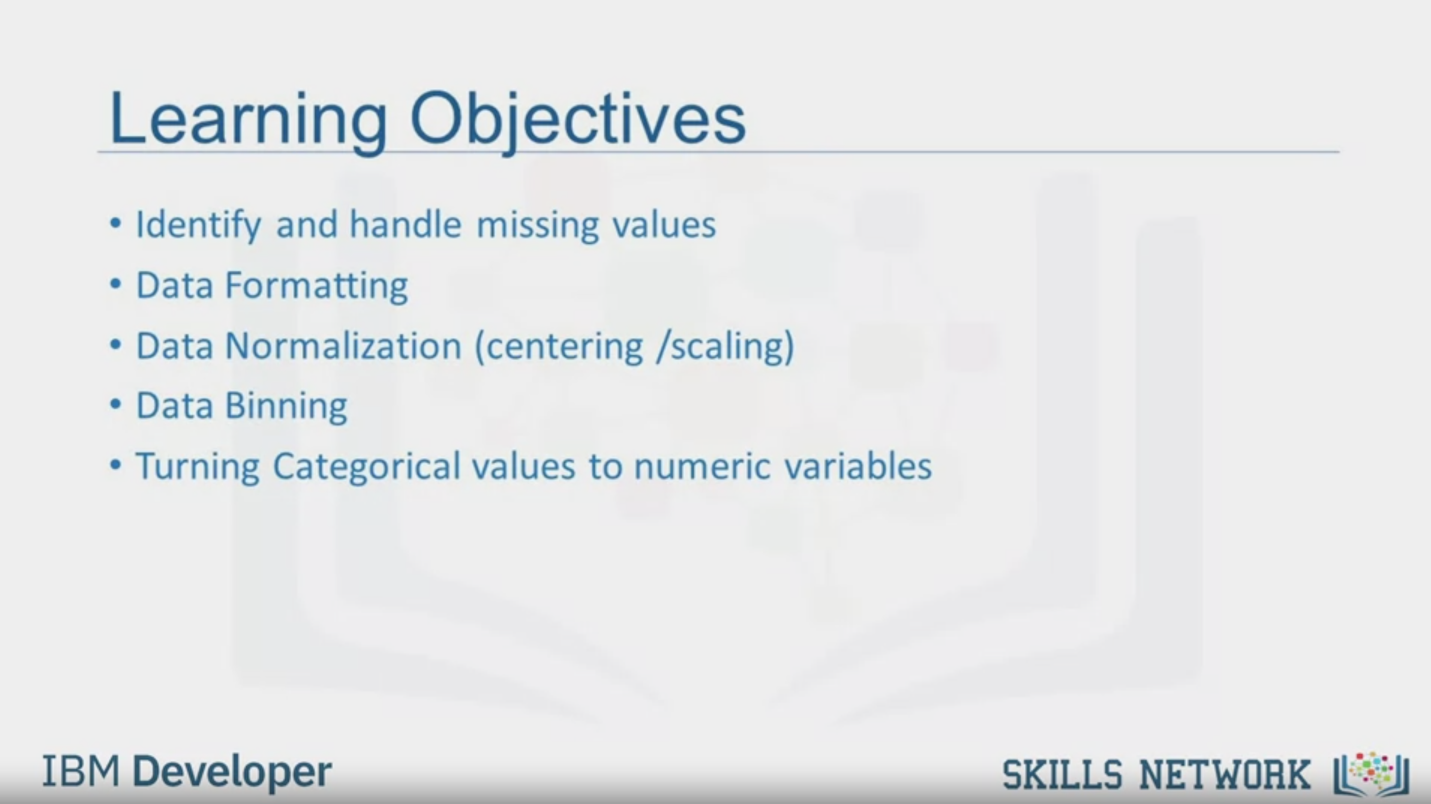
The process of converting or mapping data

from the initial “raw ” form into another format,

in order to prepare the data for further analysis.

IBM Developer

SKILLS NETWORK



Learning Objectives

• Identify and handle missing values

• Data Formatting

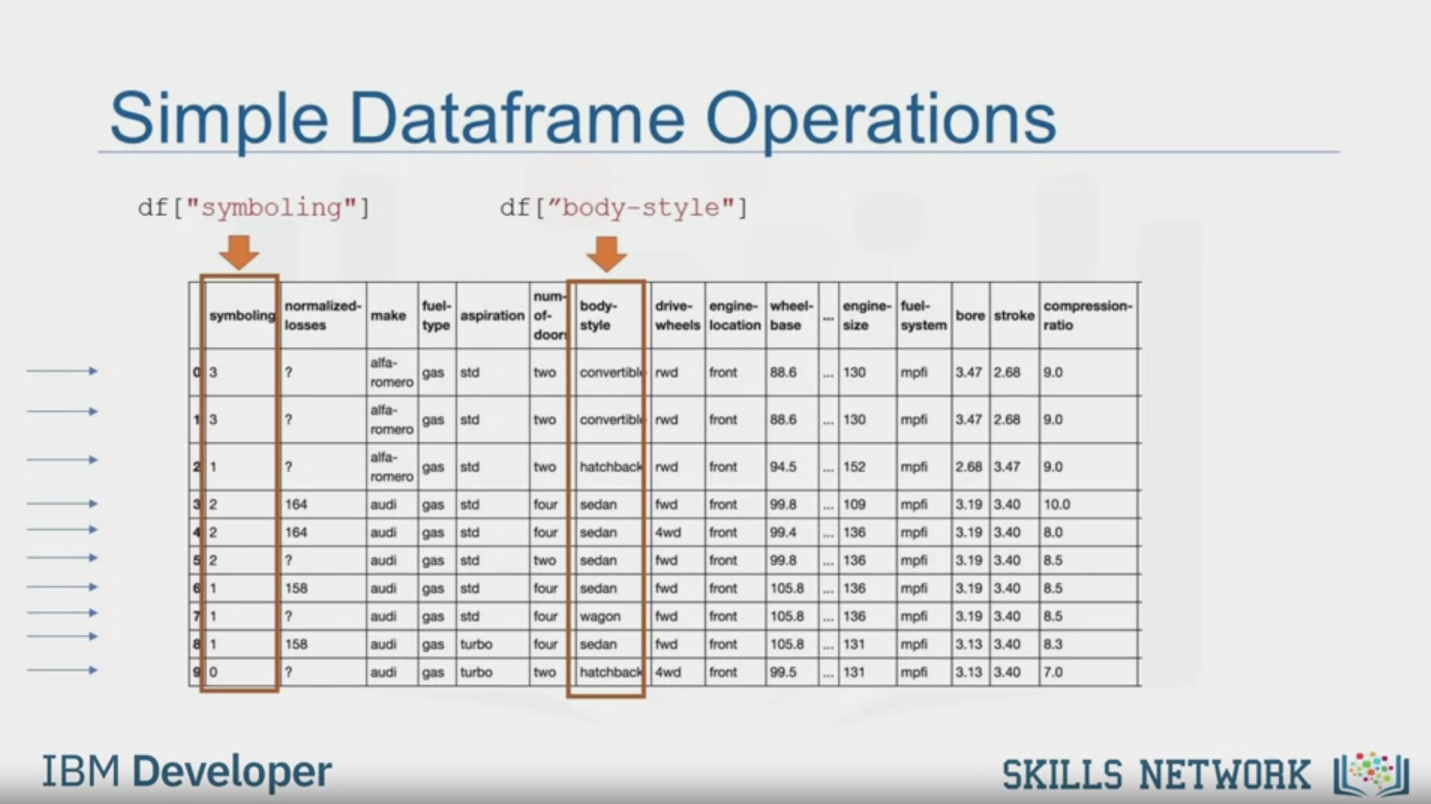
• Data Normalization (centering /scaling)

• Data Binning

• Turning Categorical values to numeric variables

IBM Developer

SKILLS NETWORK



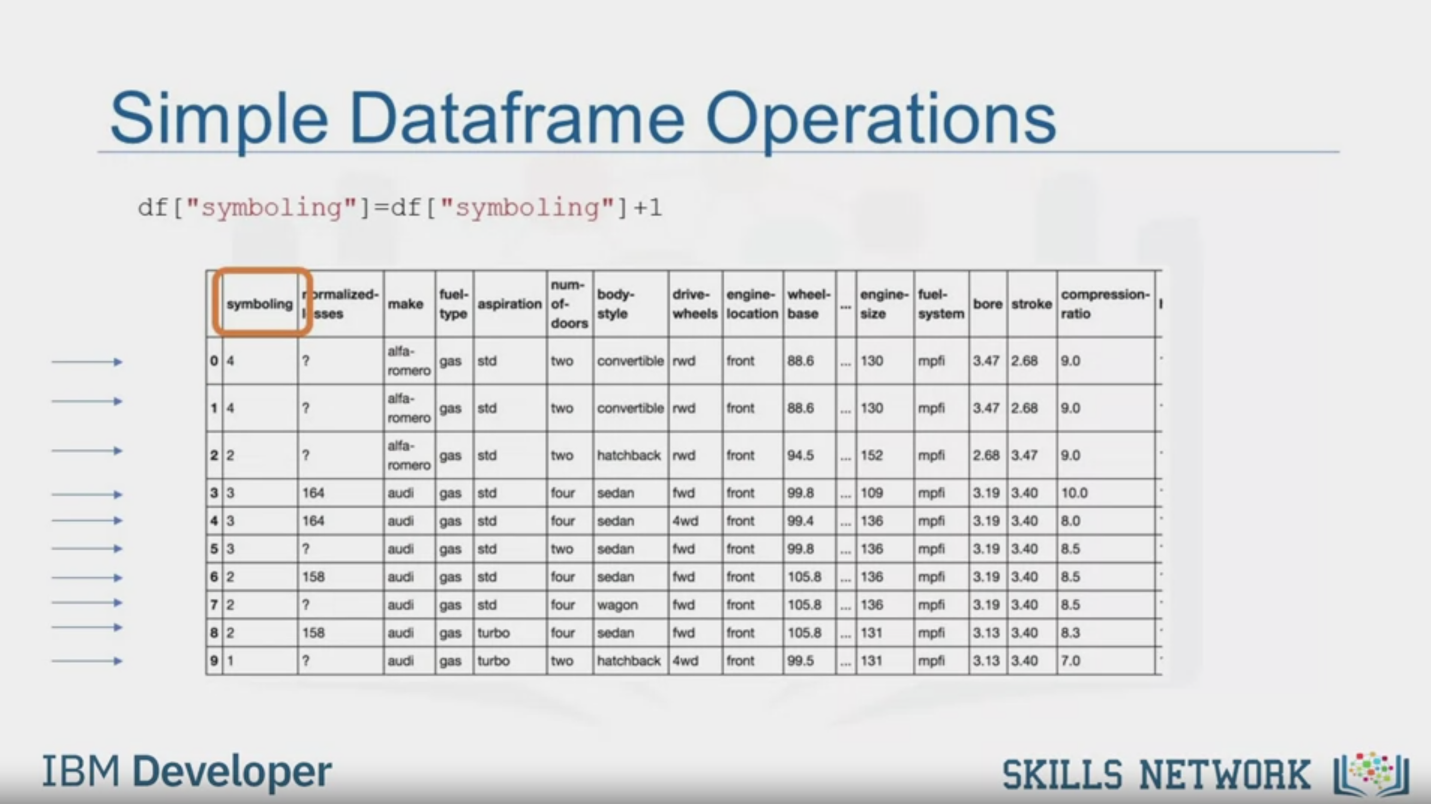
Simple Dataframe Operations

df ["symboling"]

df ["body-style"]

IBM Developer

SKILLS NETWORK



Simple Dataframe Operations

df ["symboling"]=df ["symboling" ] +1

IBM Developer

SKILLS NETWORK



SKILLS NETWORK

IBM Developer

In this video, we'll be going through

some data preprocessing techniques.

If you're unfamiliar with the term,

data preprocessing is a necessary step in data analysis.

It is the process of converting or mapping data from

one raw form into

another format to make it ready for further analysis.

Data preprocessing is often called

data cleaning or data wrangling,

and there are likely other terms.

Here are the topics that we'll be

covering in this module.

First, we'll show you how to

identify and handle missing values.

A missing value condition occurs

whenever a data entry is left empty.

Then we'll cover data formats.

Data from different sources maybe in various formats,

in different units, or in various conventions.

We will introduce some methods in

Python Pandas that can

standardize the values into the same format,

or unit, or convention.

After that, we'll cover data normalization.

Different columns of numerical data may have

very different ranges and

direct comparison is often not meaningful.

Normalization is a way to bring all data into

a similar range for more useful comparison.

Specifically, we'll focus on

the techniques of centering and scaling.

Then, we'll introduce data binning.

Binning creates bigger categories

from a set of numerical values.

It is particularly useful for

comparison between groups of data.

Lastly, we'll talk about

categorical variables and show you how to convert

categorical values into numeric variables

to make statistical modeling easier.

In Python, we usually perform operations along columns.

Each row of the column represents a sample,

I.e, a different used car in the database.

You access a column by specifying the name of the column.

For example, you can access symbolling and body style.

Each of these columns is a Panda series.

There are many ways to manipulate Dataframes in Python.

For example, you can add

a value to each entry off a column.

To add one to each symbolling entry, use this command.

This changes each value of

the Data frame column by adding one to the current value.