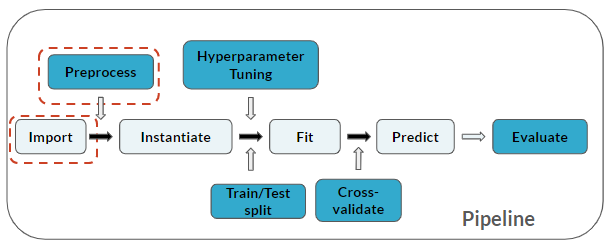
**The hunt for missing values**

Questions about processing missing values are integral to any machine learning interview. If you are provided with a dataset with missing values, not addressing them will likely skew your results and lower your model's accuracy.

In this exercise, you'll practice the first pre-processing step by finding and exploring ways to handle missing values using pandas and numpy on a customer loan dataset.

The dataset, which you'll use for many of the exercises in this course, is saved to your workspace as loan\_data.

This is where you are in the pipeline:



# Import modules

import numpy as np

import pandas as pd

# Print missing values

print(loan\_data.isna().sum())

# Fill missing values with zero

loan\_data\_filled = loan\_data.fillna(0)

# Examine 'Credit Score' before and after

print(loan\_data['Credit Score'].describe())

print(loan\_data\_filled['Credit Score'].describe())

# Simple imputation

As you saw in the last exercise, deleting data can reduce your dataset by too much. In an interview context, this can lead to biased results of your machine learning model.

A more dynamic way of handling missing values is by imputing them. There are a variety of ways of doing this in python, but in this exercise you will be using the SimpleImputer() function from the sklearn.impute module on loan\_data.

You will then use pandas and numpy to convert the imputed dataset into a DataFrame.

Note that 2 steps are now added to the pipeline, **Instantiate** and **Fit**:

