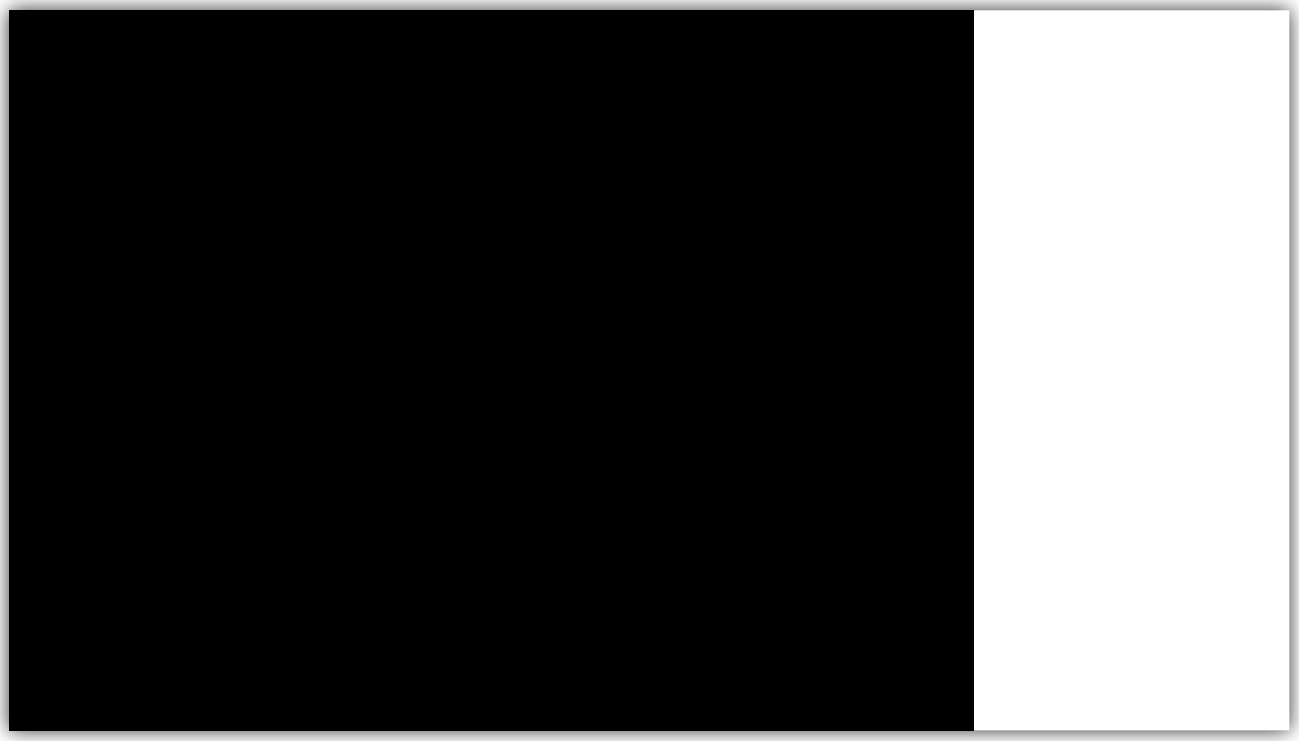
8.3.1

Overview of Data Cleaning Techniques

You need to clean the crowdfunding data before transforming it to suit the business requirements. So, you and Britta need to create a plan for how to do so.

But first, review the following video for an overview of the transform phase:



Note that the transform phase consists mostly of data cleaning. Our crowdfunding DataFrames aren't particularly messy. But sometimes, we need to clean very messy data, and each dataset presents unique challenges. So although no single right way to clean data exists, we can still have a rough technique to follow.

We determine what data cleaning techniques to use based on the state of the data. Messy data comes in three states:

- Beyond repair

- Badly damaged
- In the wrong form

Let's define each of these and explain the techniques that we might use.

Data that's **beyond repair** might have been overwritten or suffered severe data corruption during storage or transfer. (This can occur, for example, during a power loss while writing, a voltage spike, or a hard-drive failure). The worst-case scenario is data that's missing every value. All the information is lost and unrecoverable. For data beyond repair, all we can do is delete it and move on.

Data that's **badly damaged** might have good data that we can recover, but it will take time and effort to repair the damaged data. The data might be garbled, have lots of missing values, come from inconsistent sources, or exist in multiple columns. To choose the best technique, we want to consider trade-offs. Note that the best technique might not be perfect but rather the best one that's available. To repair data that's badly damaged, we can try the following techniques:

- Filling in the missing data by doing one of the following:
  - Substituting data from another source.
  - Interpolating between existing data points.
  - Extrapolating from existing data.
- Standardizing units of measure (for example, with monetary values that are stored in multiple currencies).
- Consolidating data from multiple columns.

Data that's **in the wrong form** is good but can't be used in its current form. We should usually fix this type of data. Good data that's in the wrong form might be too granular or detailed, consist of numeric data that's stored as strings, or need to be split into multiple columns (like with address data). To remedy data that's in the wrong form, we can try the following techniques:

- Reshape the data, or change the way data is organized into rows and columns (by using the `pivot()` function, for example).
- Convert data types.
- Parse text data to the correct format.
- Split columns.

The techniques for all three data states are available to us, but knowing when to perform them can feel overwhelming. No simple checklist or flowchart exists to guide us, and ultimately, that's good. The reason is that when data cleaning, we have to constantly ask ourselves what we might have missed. If we followed a rigid plan, we wouldn't ask ourselves those important questions. Data cleaning requires lots of improvising.

IMPORTANT

Document your data cleaning assumptions, decisions, and motivations for those decisions. Later decisions depend on earlier ones, which can be too much to remember. If forgotten, any assumptions that were part of an earlier decision can ruin later steps.

Needing to improvise doesn't mean that we'll be completely lost, however. We will have a strategy. Specifically, we won't try to clean the data all at once. Instead, we'll focus on one problem at a time by using an iterative process. Let's learn about that process next.

© 2020 - 2022 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.