Cleaning Data - Part I: Takeaways

Syntax

- Use **Data, Sort & Filter, Advanced, Unique records only** to identify unique records.
- Use **Data, Data Tools, Remove Duplicates** to remove duplicate records.
- `TRIM` will remove all leading and trailing spaces.
- `LEFT` and `RIGHT` will both return a specified number of characters from the left and right end of a value, respectively.
- `MID` will return a specified number of characters from a specified location in your value.
- `AVERAGE` will give you the average (or mean) of a set of numbers.

Concepts

- Data cleaning is also called data manipulation.
- Data cleaning should strive to maintain the integrity of the original dataset.
- Duplicate records are not the same thing as duplicate values that exist in distinct records.
- Extra spaces are not always visible, so you need to know how to find them.
- Separating data will often involve methods that don't include the use of delimiters.
- Inaccurate data is very common, and it is sometimes very subtle. Knowing your data is key to correcting inaccurate data.
- Numbers and text may be transposed, such as **0** and **O**, or **1** and **I**. You need to be on the lookout for these types of inaccuracies.
- Updating missing data, or imputing data, is a common task that you'll perform often as a data analyst.
- Deciding how to resolve large portions of missing data is a skill you'll develop over time as a data professional.

Resources

- LEFT function
- RIGHT function
- MID function