Cleaning and preparing data

This lesson explains the difference between clean data and raw data. It also explains why clean data is important for creating visualizations.

Raw data

Data that has been collected but hasn't had any additional processing or cleaning.

Clean data

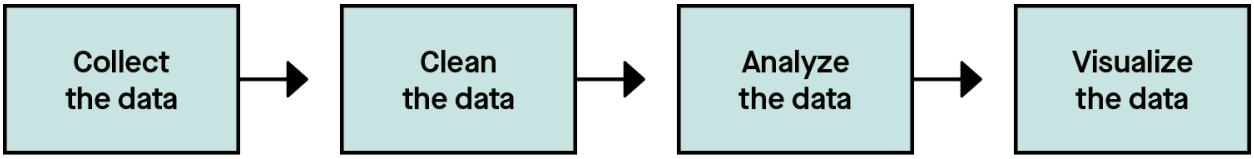
Data that has been cleansed so that it doesn't have any missing, inconsistent, or incorrect data.

Null value

A missing or blank value; also called a null.

Introduction

Recall the data analysis process that you learned in the first module of this course. As you learned, that process looks like this:



The data that you collect in the first step is raw data, which means that it hasn't undergone any additional processing. So, this data may have some issues. These issues can include the following:

- Inaccurate data.
- Inconsistent formats throughout the data.
- Missing data due to optional fields or system outages.
- Unclear field definitions that are used inconsistently by users.
- Data that is entered wrong on purpose (fraud).

Before creating visualizations based on a dataset, you need to have clean data. Leaving data raw, or unclean, can lead to a visualization that shows incorrect information. For example, if duplicate rows of data aren't removed, then a bar chart showing total amounts would show incorrectly high totals. Or, say that a column of data has numbers, but some are stored as text values and some as numeric values. Then the totals shown in a pie chart would be too low because they wouldn't account for the text values.

Inconsistent data types

One of the first things to check in a new raw dataset is the consistency and accuracy of data types. If a column has the header Temperature but the values are formatted as dates, the dataset won't make sense. Similarly, if a column has the header Number of students but contains some integers and some decimals, a visualization based on that column may be confusing. As an example, imagine that you're a contractor for a company. You must provide an invoice each week to bill the company for your services. The spreadsheet below shows your invoice for the week. What do you notice as you look through this spreadsheet?

	A	B	C
1	**Work Completed**	**Date of Service**	**Hours Worked**
2	Three sales visits	8/23/2021	8
3	Quarterly report	8/24/2021	800%
4	Slide deck preparation	8/25/2021	$8.00
5	**Total**		**24**
6			

Hopefully, you've noticed the inconsistencies in column C. The Hours Worked column has incorrect formats, including percentage and currency. Now, take a look at a cleaned version of this same data, shown below.

	A	B	C
1	**Work Completed**	**Date of Service**	**Hours Worked**
2	Three sales visits	8/23/2021	8
3	Quarterly report	8/24/2021	8
4	Slide deck preparation	8/25/2021	8
5	**Total**		**24**
6			

The Hours Worked column now has consistent data formats, with the percentages and currency changed to integers.

Duplicate rows of data

Sometimes a dataset has duplicate rows due to errors in the original data management system. If duplicate rows are left in the data, this can overstate totals and produce incorrect analyses.

Say you're trying to understand your monthly spending better. The spreadsheet below contains your expenses from last month.

	A	B
1	**Item**	**Amount**
2	Groceries	$300
3	Jeans	$75
4	Italian Restaurant	$150
5	Italian Restaurant	$60
6	Sneakers	$120
7	Hair care	$60
8	Cosmetics	$55
9	Sushi restaurant	$119
10	Rent	$1,500
11	Rent	$1,500
12	Yoga class	$35
13	Dentist appointment	$40
14	Dentist appointment	$40
15	**Total**	**$4,054**
16		

This spreadsheet shows that you've spent a total of $4,054 in the last month. But if you look carefully, you can see that Italian Restaurant, Rent, and Dentist appointment each appear twice in the list.

Notice that although Italian Restaurant shows up twice, each entry has a different amount. This indicates that these entries aren't duplicates and that you went to this restaurant two times, spending $150 the first time and $60 the second time.

If all values in a row are the same as those in another row, you can usually assume that they are duplicate rows. If only one or some of the values in a row are the same as those in another row, then they probably aren't duplicate rows.

So, only Rent and Dentist appointment have duplicate rows. You want an accurate view of what you've spent, so you want to remove these duplicates. Shown below is the same data from above, but with the duplicate rows for Rent and Dentist appointment removed.

	A	B
1	**Item**	**Amount**
2	Groceries	$300
3	Jeans	$75
4	Italian Restaurant	$150
5	Italian Restaurant	$60
6	Sneakers	$120
7	Hair care	$60
8	Cosmetics	$55
9	Sushi restaurant	$119
10	Rent	$1,500
11	Yoga class	$35
12	Dentist appointment	$40
13		
14		
15	**Total**	**$2,514**

Now the new total accurately reflects your expenditures for the last month: $2,514.

Columns of text

You may not always receive data as a spreadsheet type file. There are lots of data file types, such as text files, XML files, CSV files, and Microsoft Access Database files, to name a few. When data is obtained in a different file type, you may need to do some work to make it compatible with a spreadsheet data structure.

A common data file type is a comma-separated value file (CSV file). This file type contains data that looks like a bunch of text or numbers all within one column, and each piece of data is separated by commas. The example below shows a CSV file containing your friends' names and the holiday gifts that you've bought for each of them.

	A
1	**Name, Item1, Item2**
2	Kristy, candle, soaps
3	Nina, coffee beans, french press
4	Gavin, Wine Glasses
5	Justin, record player, record
6	Maya, planter, seeds
7	Tula, yarn, knitting book
8	Sloane, video game
9	

As you can see, this data structure makes it hard to visually separate the names from the gift items.

The image below shows the same data as above but separated out into columns.

	A	B	C
1	**Name**	**Item1**	**Item2**
2	Kristy	candle	soaps
3	Nina	coffee beans	french press
4	Gavin	Wine Glasses	
5	Justin	record player	record
6	Maya	planter	seeds
7	Tula	yarn	knitting book
8	Sloane	video game	
9			

Now your data is split out into separate columns and is much easier to look at!

Null values

If data was collected using a survey and some of the questions were optional, the data may have some blanks or null values. A null value is a blank or empty value in a set of data. The problem with null values is that they can impact the totals, averages, and other numeric calculations. If you take an average of 2, 4, and null, for example, you're potentially missing an important value that could change the average.

How are nulls dealt with? The favored way of dealing with nulls is imputing (filling in) the missing value. This is done by estimating a reasonable value for the null.

Imputing can be done in several ways. First, check if there is any other record in the data that has similar values for the other columns. If there is, it can be used as a good estimator for the missing value. Other imputation methods include finding a relevant average, median, or mode. If none of these methods seems reasonable for your null value, then the last resort is deleting the record containing the null value altogether.

The spreadsheet below shows data for different types of cardstock used in a project. As you can see, there are two null price values.

	A	B
1	**Type of cardstock**	**Price per sheet**
2	Glossy	$2
3	Matte	$2
4	Pattern - polka dots	$2
5	Pattern - stripes	
6	Gold	$1.15
7	Silver	$1.15
8	Blue	$1.15
9	Red	
10	Plain white	$0.99

The two null prices are for Pattern - stripes and Red. For Pattern - stripes, there is a similar Pattern - polka dots cardstock above it for $2. Because these are both in the same Pattern category, it's reasonable to assume that Pattern - stripes is also around $2. Next, check if there are comparable items for the Red cardstock. There are other colors—gold, silver, and blue—that are each $1.15. So, it seems reasonable to assume that the red cardstock is also $1.15.

Resolving null values requires a bit of exploration and creativity. It's best to find a reasonable estimate than to guess a value at random.

Conclusion

This lesson discussed the difference between clean data and raw data, and it explored why this is important for creating visualizations. By cleaning raw data in the ways listed above, you improve the accuracy of your visualizations.