18.2.3

Preprocessing Data With Pandas

Martha is super grateful for the Pandas refresher—it's always fun to work with such a classic dataset! Now that you are both on the same page, you start to think critically about your cryptocurrency dataset.

As mentioned, we don't know the output of the data, but that doesn't mean we shouldn't think about our data or that we should carelessly plug it into a model.

Let's take a look at how we should start our data processing by loading in the **shopping_data.csv** (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/module_18/shopping_data.csv)

```
# Data loading
file_path = "Resources/shopping_data.csv"
df_shopping = pd.read_csv(file_path, encoding="ISO-8859-1")
df_shopping.head(5)
```

	CustomerID	Card Member	Age	Annual Income	Spending Score (1–100)
0	1	Yes	19.0	15000	39.0
1	2	Yes	21.0	15000	81.0
2	3	No	20.0	16000	6.0
3	4	No	23.0	16000	77.0
4	5	No	31.0	17000	40.0

Questions for Data Preparation

Unsupervised learning doesn't have a clear outcome or target variable like supervised learning, but it is used to find patterns. By properly preparing the data, we can select features that help us find patterns or groups.

Before we begin, consider these questions:

- What knowledge do we hope to glean from running an unsupervised learning model on this dataset?
- What data is available? What type? What is missing? What can be removed?
- Is the data in a format that can be passed into an unsupervised learning model?
- Can I quickly hand off this data for others to use?

Let's address the first question on our list:

What knowledge do we hope to glean from running an unsupervised learning model on this dataset?

It's a shopping dataset, so we can group together shoppers based on spending habits.