The Dataset

Learn about the retail dataset used for the project.

Chapter Goals:

- Learn about the retail dataset used in this case study
- Read the separate data files that comprise the dataset

A. Starting off the project

Let's say you're a machine learning engineer at a large retail corporation, and your supervisor just gave you this dataset and said,

"I want a system that can make future sales predictions for these 45 stores. We want to know whether they'll make enough money to justify keeping them. Here's a dataset containing the past sales of these stores."

This is a pretty short and vague description of the project, which is normally the type of description you'd get from a manager or supervisor. Luckily, we can learn more about the project by looking through the dataset.

Industry data usually comes in CSV files, XLSX spreadsheets, JSON data files, or can be accessed from a database using SQL. In this case, our dataset comes in three CSV files: *weekly_sales.csv*, *features.csv*, and *stores.csv*. We'll use the pandas library's `pd.read_csv` function to read each CSV file into a DataFrame.

For more on pandas and data processing, check out the Machine Learning for Software Engineers course on Educative.

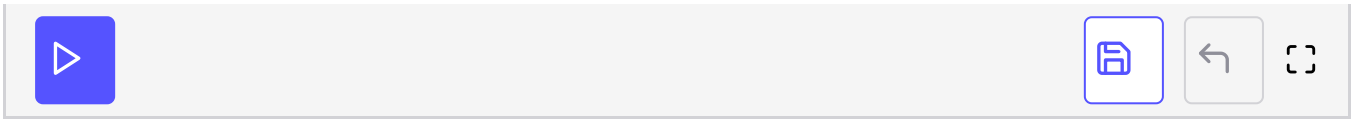
B. Understanding the dataset

Your supervisor gave you some basic details about the dataset. The *weekly_sales.csv* file contains rows detailing the sales (in dollars) for the departments of each store in a given week. We use this file to train a machine learning model to make future weekly sales predictions for each store's department.

```
import pandas as pd

train_df = pd.read_csv('weekly_sales.csv')

print(train_df)
```



DataFrame (train_df) containing data from weekly_sales.csv. The DataFrame contains 421570 rows.

After taking a look at the data using pandas, you confirm that the *weekly_sales.csv* file does indeed match your supervisor's description. There's also an additional column called `'Holiday'`, which is `True` if the row's week has a holiday, otherwise it's `False`.

The *features.csv* file contains potentially useful features, with values given on a weekly basis for each store. These features include a given week's national unemployment rate and the temperature of the region that the store is located in. The *stores.csv* file contains information about each of the 45 stores, specifically the type of store and the size of the store.

We'll take a deeper look at the features and stores CSV files in later chapters.

Time to Code!

In this chapter you'll be completing the `read_dataframes` function, which returns a DataFrame for each of the CSV files in the dataset: *weekly_sales.csv*, *features.csv*, and *stores.csv*.

We can use pandas to read the CSV data into DataFrame objects, then return the three DataFrames.

Set `train_df` equal to `pd.read_csv` applied with `'weekly_sales.csv'`.

Set `features_df` equal to `pd.read_csv` applied with `'features.csv'`.

Set `stores_df` equal to `pd.read_csv` applied with `'stores.csv'`.

Return a tuple containing the three created DataFrames, `train_df`, `features_df`, and `stores_df` (in that order).

```
import pandas as pd

# Read the CSV data files into DataFrames

def read_dataframes():
    # CODE HERE
    pass
```