Introduction

In this section of the course you will be performing data processing on the final dataset from the **Preliminary Data Analysis** section. Specifically, you'll be building the input pipeline for training and evaluating the machine learning model.

A. Additional data processing

In the **Preliminary Data Analysis** section, we performed data analysis on the retail dataset and concluded that there is a strong enough correlation from the dataset's features to predict weekly store sales.

The files in that chapter were small enough to use relatively simple techniques. If the files were larger and more resource intensive, we would have used the techniques laid out in the **Efficient Data Processing Techniques** section.

After informing the project supervisor that the prediction task is viable, we begin working on the machine learning model.

However, before writing any actual machine learning code, we know that we need to continue processing the data to create an efficient *input pipeline*. The input pipeline represents how the data will be passed into the model for each step of training or evaluation. Since training the model requires thousands of steps, it is important that the input pipeline is as efficient as possible.

The final dataset we created was stored in a pandas DataFrame. Since the DataFrame is not the most efficient data storage for the input pipeline, we'll need to perform additional processing to create a more efficient solution.