Categorical Data

Learn about categorical data and how it is used in a dataset.

Chapter Goals:

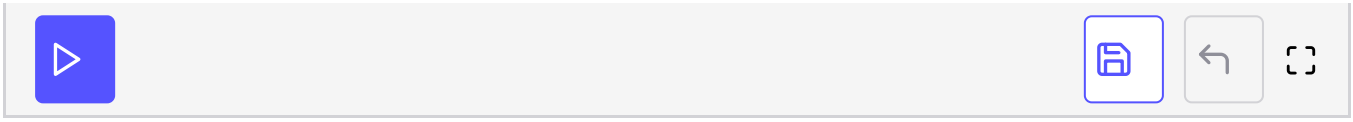
- Analyze the categorical features in the dataset

A. The dataset format

When using the dataset to train a machine learning model, each feature needs to be an integer, float, or string type. The float data is the numeric data, i.e. the data that can be quantified and analyzed using operations like mean or standard deviation. The string data is categorical, meaning that each string represents some unique category for the feature. Integer data can be either numeric (e.g. kilometer distance) or categorical (e.g. year of birth).

In the final dataset we're using, the categorical features are the `'Store'`, `'Type'`, and `'Dept'` features. Since we already know there are 45 stores, labeled from 1 to 45, we just need to investigate the `'Type'` and `'Dept'` features.

```
print(final_dataset['Type'].unique())
print(final_dataset['Dept'].unique())
```



There are only three categories of stores shown in the `'Type'` feature: `'A'`, `'B'`, and `'C'`. There are 81 store departments, and each is a positive integer less than 100. We'll discuss how to process and use categorical and numeric features in a machine learning model later in this course.

Time to Code!

When using the dataset for training a machine learning model, we want the features to either be integer, float, or string type. Currently the only feature in

`final_dataset` that is not one of those types is `'IsHoliday'`. Therefore, the

coding exercise for this chapter is to convert the `'IsHoliday'` boolean values to binary integers.

Cast the `'IsHoliday'` feature of `final_dataset` to `int` type.

```
import pandas as pd

# Update IsHoliday values
# CODE HERE
```