Missing Data

Learn how to merge feature datasets find missing data in features.

Chapter Goals:

- Merge the two DataFrames containing the feature data
- Learn how to identify features with missing data

A. Merging the features

Both the `features_df` and `stores_df` DataFrames contain feature data, i.e. data related to the stores or weeks that correspond to the rows in `train_df`. Remember each feature is a column of the DataFrame, and that each row in a given column is an entry for that feature. Let's take a look at the specific features contained in each DataFrame.

```
general_features = features_df.columns

print(general_features)
print('General Features: {}\n'.format(general_features.tolist()))

store_features = stores_df.columns
print('Store Features: {}'.format(store_features.tolist()))
```

The code above shows the features (i.e. the columns) of the `features_df` and `stores_df` DataFrames. The `tolist` function converts the features from an `Index` to a list.

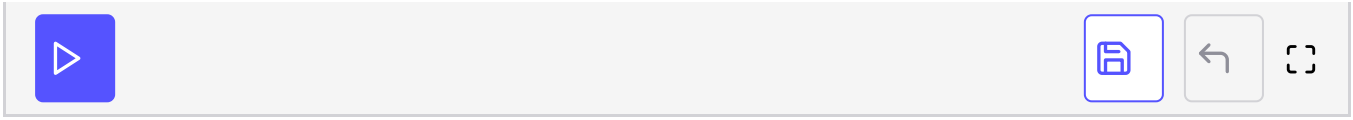
You'll notice that both these DataFrames share the `'Store'` feature, which is just the ID of the store for a given row in the DataFrame. Since both these DataFrames contain useful data, we can make things easier for ourselves by merging the two DataFrames into one.

We do this by using the `merge` function and merging the DataFrames based on

the `'Store'` feature.

```
merged_features = features_df.merge(stores_df, on='Store')

print(merged_features)
```



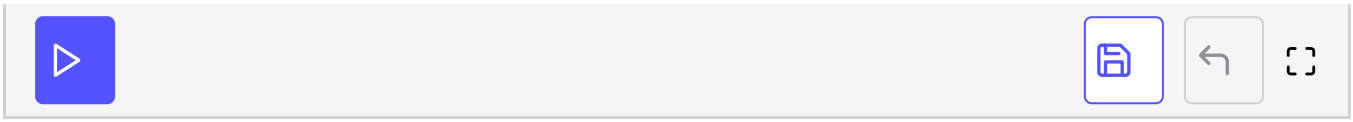
The code above merges the two DataFrames. The new DataFrame (`merged_features`) contains all the features from `features_df` and `stores_df`. It has a total of 8190 rows.

B. Finding missing data

Using the newly merged DataFrame, we can figure out which features contain missing data. This is a crucial step in the data analysis, since we need to perform data processing on any features that have missing row values (i.e. only some of the rows will have entries for that feature).

The pandas library will represent missing values with an `NA` in the DataFrame. We use the `pd.isna` function combined with the `any` function to check which columns contain missing values.

```
na_values = pd.isna(merged_features) # Boolean DataFrame
na_features = na_values.any() # Boolean Series
print(na_features)
```



The `'CPI'` and `'Unemployment'` features contain missing values, along with each `'Markdown'` feature. We'll discuss how to handle these missing values in the next chapter.