Programming Assignment: Cafe Sales Data Cleaning

This assignment focuses on data cleaning and preprocessing techniques using pandas and numpy. You will work with a dirty dataset of cafe sales, which contains various common data quality issues. Your task is to clean the dataset and prepare it for analysis.

Assignment Description

Your task is to write a Python script that performs the following steps:

1. **Load the Dataset:** The input will be a CSV file named dirty_cafe_sales.csv. You must use the pandas library to load this file into a DataFrame.
2. **Data Type Correction:** Inspect the data types of each column and convert them to the most appropriate types. For example, numerical columns should not be of object type, and dates should be in a datetime format.
3. **Handle Missing and Invalid Values:** The dataset contains missing values (e.g., NaN, empty cells) and invalid entries (e.g., "ERROR", "UNKNOWN"). Your task is to:
    - Identify and handle these values in a suitable manner. For numerical columns, consider imputation with the mean or median, BUT if those options are not appropriate for any columns, drop the rows with missing values. For categorical columns, you might replace them with the mode or a specific category like "Unknown", or again, drop them if it is appropriate.
    - Data type correction for some columns might not work unless the invalid values are handled.
4. **Data Consistency and Integrity:**
    - **Item Names:** Some item names are missing or invalid. Based on the Price Per Unit, you should be able to find out the correct item name for some of the missing values.
    - **Price and Total Spent:** There are missing Price Per Unit and Total Spent values. You should calculate these missing values based on the relationship: Total Spent = Quantity * Price Per Unit.
5. **New Feature:** Create a new column called ‘season’ that specifies which season of the year the sale was made in (Winter/Spring/Summer/Fall). You can derive this via the date column.
6. **Output Cleaned Data:** Save the cleaned DataFrame to a new CSV file named cleaned_cafe_sales.csv.

Hints

- Use pandas functions like to_numeric, to_datetime, fillna, and replace for data cleaning.
- Pay close attention to the data characteristics described on the Kaggle dataset page.
- Document your cleaning decisions and the reasons for them in your code using markdown cells in jupyter.

Good luck!