CCA-2: Data Load and Initial Cleaning

Overview

This section outlines how to load the dataset into the analysis environment and perform initial data cleaning and formatting. These steps prepare the data for further processing and analysis.

Objectives

- Load the raw dataset from the specified source.
- Perform initial data cleaning and formatting.

Steps

1. Load the Dataset

The first task is to load the dataset from its source using dynamic paths. This ensures the code is portable and can be executed in any environment.

Code Snippet

```
import os
import pandas as pd
import json

# Helper function to find the project root directory
def find_project_root(filename='config.json'):
    current_dir = os.path.abspath('')
    while True:
        if filename in os.listdir(current_dir):
            return current_dir
        parent_dir = os.path.dirname(current_dir)
        if parent_dir == current_dir:
            raise FileNotFoundError(f"{filename} not found in any parent directories.")
        current_dir = parent_dir

# Find the project root directory
root_dir = find_project_root()

```

```
# Load configuration
config_path = os.path.join(root_dir, 'config.json')
with open(config_path, 'r') as f:
    config = json.load(f)

# Load the dataset using the dynamic path from the config file
raw_data_path = os.path.join(root_dir, config['raw_data_path'])
df = pd.read_csv(raw_data_path)
print("Dataset loaded successfully.")
print(df.head())

```

Explanation

- **Dynamic Path Configuration**: The `find_project_root` function ensures the script can dynamically find the project's root directory, which is crucial for code portability.
- **Loading the Dataset**: The dataset is loaded using the path specified in the `config.json` file, making it easy to locate and load the data regardless of the environment.

2. Initial Data Cleaning and Formatting

After loading the data, the next step is to clean and format it. This involves handling missing values and encoding categorical variables.

Code Snippet

```
from handle_missing_and_encode import handle_missing_and_encode

# Handle missing data and encode categorical variables
df_cleaned = handle_missing_and_encode(df)
print("Data cleaned successfully.")
print(df_cleaned.head())
```

Explanation

- **Handling Missing Data**: The `handle_missing_and_encode` function fills in missing values in the dataset. This step is essential to avoid inaccuracies in analysis and model predictions.
- **Encoding Categorical Variables**: Categorical variables are converted into a numerical format suitable for machine learning algorithms.

Importance of Initial Data Cleaning

- **Ensuring Data Quality**: Initial data cleaning maintains high data quality, which is essential for accurate analysis and modeling.
- **Handling Missing Values**: Properly handling missing values prevents biases and errors in the analysis.
- **Encoding Categorical Variables**: Encoding ensures that categorical variables can be used in machine learning models, which typically require numerical input.

Results

After completing the data load and initial cleaning steps, the following results were achieved:

- **Dataset Loaded**: The raw dataset was successfully loaded into the analysis environment.
- **Data Cleaned**: Initial cleaning was performed, handling missing values and encoding categorical variables.
- **Data Saved**: The cleaned dataset was saved to the specified path in the configuration file.

Next Steps

With the data now loaded and cleaned, the next steps involve:

Exploratory Data Analysis (EDA)

- Conduct detailed exploratory data analysis to gain insights into the dataset.
- Visualize data distribution and identify patterns or anomalies.

Feature Engineering

- Create new features to enhance the predictive power of models.
- Transform existing features to better represent the underlying data.

Data Splitting

- Split the cleaned dataset into training and testing sets for model evaluation.

Model Training and Evaluation

- Train predictive models using the training set.
- Evaluate models using the testing set to assess performance.

Reporting

- Generate reports and visualizations to present findings from the analysis and modeling.

By following these steps, we ensure a comprehensive data pipeline that maintains data quality and supports effective data analysis and modeling.

Conclusion

Loading and performing initial cleaning of the dataset is a crucial step in the data engineering process. By ensuring that the data is free from initial errors and inconsistencies, we lay a solid foundation for accurate and effective data analysis and modeling. This documentation provides a clear and detailed guide on how to perform these tasks, ensuring that the process is reproducible and understandable.