Data Cleaning Project with Python

Project Overview

A comprehensive data cleaning pipeline that processes raw CSV files, performs data profiling and cleaning operations, and outputs cleaned datasets with detailed logging and documentation.

Features

- **Data Profiling**: Comprehensive analysis and statistics generation
- **Cleaning Pipeline**: Configurable data cleaning operations
- **Multiple Approaches**: Script, Functional, or Object-Oriented implementations
- **Comprehensive Logging**: Track all operations with multiple log levels
- **Configuration Management**: JSON-based configuration system
- **Documentation**: Complete docstrings and README

Project Structure

```
data-cleaning-project/
├── src/
├── data/
├── logs/
├── requirements.txt
└── README.md
```

Installation

```
git clone <repository-url>
cd data-cleaning-project
pip install -r requirements.txt
```

Usage

```
from src.data_cleaner import DataCleaner

cleaner = DataCleaner('config/cleaning_config.json')
cleaned_data = cleaner.clean_data('input.csv')
```

Configuration

Edit `config/cleaning_config.json` to customize cleaning operations:

- Drop null values

- Remove duplicates
- Handle outliers
- Data type corrections