Automated Data Cleaning & Processing Pipeline

1. Define Scope

Objective: Develop an automated pipeline to clean and preprocess datasets.
Target Users: Data analysts, data scientists, and businesses working with raw data.
Key Tasks:
- Remove missing values
- Standardize column names and data formats
- Remove duplicates
- Handle outliers
- Save the cleaned data for analysis

2. Data Collection

We'll use the 'National Drug Code Directory' dataset, which contains drug listings with potential issues like missing values and duplicates. This dataset is publicly available and suitable for practicing data cleaning techniques.

3. Data Cleaning & Preprocessing Pipeline

Below is the Python code to create an automated data cleaning pipeline using Pandas and NumPy:

```
import pandas as pd
import numpy as np

def load_data(file_path):
    return pd.read_csv(file_path)

def standardize_column_names(df):
    df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')
    return df

def handle_missing_values(df, strategy='mean', fill_value=None):
    if strategy == 'drop':
        df = df.dropna()
    elif strategy == 'fill':
        df = df.fillna(fill_value)
    elif strategy == 'mean':
        df = df.fillna(df.mean(numeric_only=True))
    elif strategy == 'median':
        df = df.fillna(df.median(numeric_only=True))
    elif strategy == 'mode':
```

```
        df = df.fillna(df.mode().iloc[0])
    return df

def remove_duplicates(df):
    df = df.drop_duplicates()
    return df

def handle_outliers(df, z_thresh=3):
    numeric_cols = df.select_dtypes(include=[np.number]).columns
    z_scores = np.abs((df[numeric_cols] - df[numeric_cols].mean()) / df[numeric_cols].std())
    df = df[(z_scores < z_thresh).all(axis=1)]
    return df

def save_clean_data(df, output_path):
    df.to_csv(output_path, index=False)

def data_cleaning_pipeline(file_path, output_path, missing_value_strategy='mean',
fill_value=None, z_thresh=3):
    df = load_data(file_path)
    df = standardize_column_names(df)
    df = handle_missing_values(df, strategy=missing_value_strategy, fill_value=fill_value)
    df = remove_duplicates(df)
    df = handle_outliers(df, z_thresh=z_thresh)
    save_clean_data(df, output_path)
    return df
```

4. Findings & Insights

- Missing Values: Handled using the specified strategy, ensuring no missing data remains.
- Duplicates: Removed duplicate entries, resulting in a unique set of records.
- Outliers: Identified and removed outliers based on the z-score threshold, ensuring data consistency.

5. Conclusion & Recommendations

Conclusion: An automated data cleaning pipeline effectively preprocesses raw datasets, addressing common issues such as missing values, duplicates, and outliers.
Recommendations:
- Integrate the pipeline into data ingestion processes to maintain data quality.
- Customize the pipeline functions as needed to handle dataset-specific challenges.
- Regularly update and test the pipeline to adapt to evolving data quality requirements.