A PROJECT REPORT ON

"Python Data Preprocessing Library (Publishable on PyPI)"

SUBMITTED BY

NAME	:	MHD ALHABEB ALSHALAH, Ahmed Emad Elsayed Mohamed Abdelfattah, MOHAMED RAGAB ABDELFATTAH ABDELFADEEL
ID	:	2221251360, 230222101, 2221251356
COURSE	:	SEN22250E Programming for Data Engineering

PyTHub

Installation

To install PyTHub, use the following command:

‘pip install pythub’

Features

PyTHub includes the following modules and functionalities:

-MissingValueHandler: Handle missing values by replacement or deletion.

- replace_value(df, value)
- impute_mean(df, columns)
- impute_median(df, columns)
- impute_constant(df, columns, value)
- delete_missing(df, columns)
- get_rows_with_missing_data(df, column)

-OutlierHandler: Detect and handle outliers using the IQR method.

- iqr_outliers(df, column, threshold=1.5)

-Scaler: Standardize and normalize data.

- standard_scale(df, columns)
- minmax_scale(df, columns)

-TextCleaner: Perform string manipulation and text cleaning.

- remove_stopwords(text)
- to_lowercase(text)
- remove_punctuation(text)
- lemmatize(text)
- clean_text(text)
- clean_columns(df, columns)

-FeatureEngineer: Create new features.

- create_feature(df, column, func)

-DataTypeConverter: Convert data types.

- to_numeric(df, columns)
- to_categorical(df, columns)

-CategoricalEncoder: Encode categorical data.

- one_hot_encode(df, columns)
- label_encode(df, columns)

-DateTimeHandler: Handle date and time data.

- to_datetime(df, columns)
- extract_date_parts(df, column)

-DataFrameLoader: Read data from CSV files.

- read_csv(file_path, **kwargs)

-DataSorter: Sort DataFrame rows.

- sort_by_row(df, column, ascending=True)

-DataFrameModifier: Modify DataFrames.

- delete_column(df, column)

Technical Requirements:

Python Programming: Utilize Python for library development.

Object-Oriented Programming (OOP): Implement functions using classes and methods.

Data Manipulation Libraries: Leverage pandas and NumPy for data handling.

Unit Testing: Implement unit tests to ensure code functionality and robustness.

Documentation: Provide clear documentation for each function with usage examples.

PyPI Packaging: Package the library for publishing on PyPI.

Description of each method:

CategoricalEncoder

ChangingValue

DataFrameLoader

DataFrameModifier

DataSorter

DataTypeConverter

DateTimeHandler

FeatureEngineer

MissingValueHandler

OutlierHandler

TextCleaner