

SOFTWARE REQUIREMENTS SPECIFICATION (SRS)

1. System Overview

SmartClean is a modular Python package designed to operate on tabular data structures using pandas DataFrames.

The system consists of several core modules responsible for specific cleaning tasks.

2. System Architecture

The architecture follows a **modular pipeline design**.

```
graph TD; A[Data Input] --> B[Dataset Profiling]; B --> C[Cleaning Modules]; C --> D[Pipeline Engine]; D --> E[Report Generator];
```

3. Module Architecture

Core Modules

IO Module

Responsible for loading datasets.

Functions:

```
read_csv()  
read_excel()  
read_json()
```

Profiling Module

Responsible for dataset inspection.

Responsibilities:

- detect missing values
- detect duplicates
- detect data types
- detect potential outliers

Missing Data Module

Handles missing values.

Functions:

```
fill_mean()  
fill_median()  
fill_mode()  
fill_auto()
```

Duplicate Module

Handles duplicate rows.

Functions:

```
detect_duplicates()  
remove_duplicates()
```

Type Correction Module

Responsible for correcting incorrect data types.

Functions:

```
convert_numeric()  
convert_datetime()  
convert_boolean()
```

Text Cleaning Module

Handles string normalization.

Functions:

```
strip_whitespace()  
normalize_case()  
remove_special_chars()
```

Outlier Module

Responsible for identifying extreme values.

Methods:

```
IQR detection  
Z-score detection
```

Column Module

Handles column name normalization.

Functions:

```
snake_case_columns()  
clean_column_names()
```

Pipeline Module

Orchestrates cleaning steps.

Function:

```
auto_clean()
```

Reporting Module

Generates summaries.

Functions:

```
generate_report()
```

4. Functional Requirements

FR1 – Dataset Loading

The system must load datasets from CSV, Excel, JSON, and pandas DataFrames.

FR2 – Dataset Profiling

The system must analyze datasets and report data quality issues.

FR3 – Missing Value Handling

The system must detect and resolve missing values.

FR4 – Duplicate Removal

The system must detect and remove duplicate rows.

FR5 – Type Correction

The system must convert incorrect data types.

FR6 – Text Normalization

The system must clean textual data fields.

FR7 – Outlier Detection

The system must detect outliers using statistical methods.

FR8 – Automatic Cleaning

The system must support an automatic cleaning pipeline.

FR9 – Reporting

The system must generate cleaning reports.

5. Non-Functional Requirements

Performance

The system should handle datasets with up to **1 million rows** efficiently.

Usability

API must remain simple and intuitive.

Reliability

Cleaning operations must not corrupt original datasets.

Extensibility

Architecture should support future modules.

6. Dependencies

Primary dependencies:

- pandas
 - numpy
 - scipy (optional)
-

7. Version Roadmap

Version 0.1

Basic cleaning functions.

Version 0.5

Full cleaning pipeline.

Version 1.0

Stable release with reporting and profiling.

8. Risks

Potential challenges:

- automated cleaning decisions may not always be correct
- handling diverse datasets reliably
- performance on large datasets