Data Cleaning in Python

Techniques and Functions for Effective Data Preparation

Introduction to Data Cleaning

Definition: The process of identifying and correcting (or removing) inaccurate records from a dataset.

Importance: Ensures data quality, improves analysis accuracy, and enhances decision-making.

Common Data Quality Issues

- Missing values
- Duplicates
- Incorrect data types
- Outliers
- Inconsistent formatting

Python Libraries for Data Cleaning

- **Pandas**: A powerful data manipulation library.
- **NumPy**: Useful for numerical operations.
- **OpenPyXL**: For Excel file manipulation.
- **Regex (re)**: For string pattern matching and manipulation.

Handling Missing Values

Functions:

- **isnull():** Identify missing values.
- **dropna() :** Remove missing values.
- **fillna() :** Fill missing values with a specified value or method.

```python
#python
import pandas as pd

#Example
df = pd.DataFrame({'A': [1, 2, None], 'B': [4, None, 6]})
df.dropna()  # Remove rows with missing values
df.fillna(0)  # Replace missing values with 0
```

Removing Duplicates

Functions:

- **drop_duplicates()** : Remove duplicate rows from a DataFrame.

```
df.drop_duplicates()
```

Correcting Data Types

Functions:

- **astype():** Convert data types of columns.

```
df['A'] = df['A'].astype(int)  # Convert column A to integer
```

Handling Outliers

- Methods:
  - Z-score method
  - IQR (Interquartile Range) method

- Example:Using boolean indexing to filter out outliers.

```
# Example using IQR
Q1 = df['A'].quantile(0.25)
Q3 = df['A'].quantile(0.75)
IQR = Q3 - Q1
df_filtered = df[(df['A'] >= (Q1 - 1.5 * IQR)) & (df['A'] <= (Q3 + 1.5 * IQR))]
```

Standardizing Formats

Functions:

- **str.lower()** , **str.upper()** : Change string case.
- **str.replace()** : Replace substrings.

```
df['B'] = df['B'].str.lower()  # Convert to lowercase
```

Conclusion

- Data cleaning is a crucial step in data analysis.
- Python provides powerful tools and functions to facilitate the cleaning process.
- Clean data leads to better insights and decision-making.