Data Cleaning

A Quick Revision

Prasad Deshmukh

Data Cleaning

- Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset. It involves several steps to ensure that the data is reliable and ready for analysis.

Identify Missing Values

- Check for any missing data points in the dataset, and decide whether to fill in the missing values or remove the corresponding entries.

```
# Check for missing values
df.isnull().sum()

# Fill missing values
df.fillna(value, inplace=True)

# Remove entries with missing values
df.dropna(inplace=True)
```

Remove Duplicates

- Detect and eliminate duplicate records from the dataset to prevent redundancy and skewed analysis.

```
# Remove duplicates
df.drop_duplicates(inplace=True)
```

Handle Inconsistent Data Formats

- Standardize data formats such as dates, addresses, and phone numbers to ensure consistency and accurate comparisons.

```
# Standardize date format
df['date'] = pd.to_datetime(df['date'], format='%Y-%m-%d')

# Normalize phone numbers
df['phone'] = df['phone'].str.replace('-', '')

# Normalize addresses
df['address'] = df['address'].str.title()
```

Correct Data Errors

- Identify and rectify any errors or outliers that may affect the quality of the dataset.

```
# Identify and correct errors
df.loc[df['age'] < 0, 'age'] = 0
```

Handle Outliers

- Determine if outliers are genuine data points or errors, and decide whether to remove them or treat them separately in the analysis.

```
# Identify and remove outliers using z-score
from scipy import stats
z_scores = stats.zscore(df['value'])
df = df[(z_scores < 3)]
```

Resolve Inconsistencies

- Deal with inconsistent data entries or conflicting information by applying data transformation or merging techniques.

```
# Merge inconsistent categories
df.replace({'category': {'Cat': 'Category', 'Cate': 'Category'}}, inplace=True)
```

Validate And Verify Data

- Cross-check the data against reliable sources or business rules to validate its accuracy and integrity.

```
# Cross-check data against a reference source
df = df.merge(reference_df, on='id', how='inner')
```

In conclusion, data cleaning plays a crucial role in ensuring the quality, reliability, and integrity of data used in analysis and decision-making processes. By identifying and rectifying errors, inconsistencies, and outliers, data cleaning enhances the accuracy of insights derived from the data, enabling more informed and reliable outcomes in various domains such as business, research, and data-driven applications. It is a critical step in the data science workflow that contributes to the overall effectiveness and credibility of data analysis.

Prasad Deshmukh

THANK YOU

Prasad Deshmukh