

EXPLORATORY DATA SCIENCE

Zhabiyan Surya Wiradenish

DATA SCIENCE

IMPORT & LOAD DATA



This code imports the Wine Quality dataset using the Pandas library and displays 1,599 wine samples containing various physicochemical properties such as acidity, residual sugar, pH, and alcohol content. Each sample is also labeled with a quality score that reflects the wine's overall rating. This dataset is well-suited for data profiling, pattern discovery, and predictive modeling using regression or classification techniques.

Kode ini menggunakan library Pandas untuk memeriksa struktur dari dataset Wine Quality yang disimpan dalam variabel df. Metode df.info() memberikan ringkasan singkat mengenai dataset, termasuk:

- Jumlah total entri (1599 baris, diindeks dari o hingga 1598).
- Jumlah total kolom (12 kolom secara keseluruhan).
- Nama kolom, jumlah nilai yang tidak kosong (nonnull), serta tipe data dari setiap kolom.
- Penggunaan memori dari DataFrame (sekitar 150.0 KB).

```
[10] df.info()
    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 1599 entries, 0 to 1598
    Data columns (total 12 columns):
                                Non-Null Count Dtype
          fixed acidity
                                1599 non-null
                                                float64
          volatile acidity
                                1599 non-null
                                                float64
          citric acid
                                1599 non-null
                                                float64
          residual sugar
                                1599 non-null
                                                float64
          chlorides
                                                float64
                                1599 non-null
          free sulfur dioxide
                              1599 non-null
                                                float64
         total sulfur dioxide 1599 non-null
                                                float64
          density
                                1599 non-null
                                                float64
                                                float64
                                1599 non-null
          sulphates
                                1599 non-null
                                                float64
         alcohol
                                1599 non-null
                                                float64
```

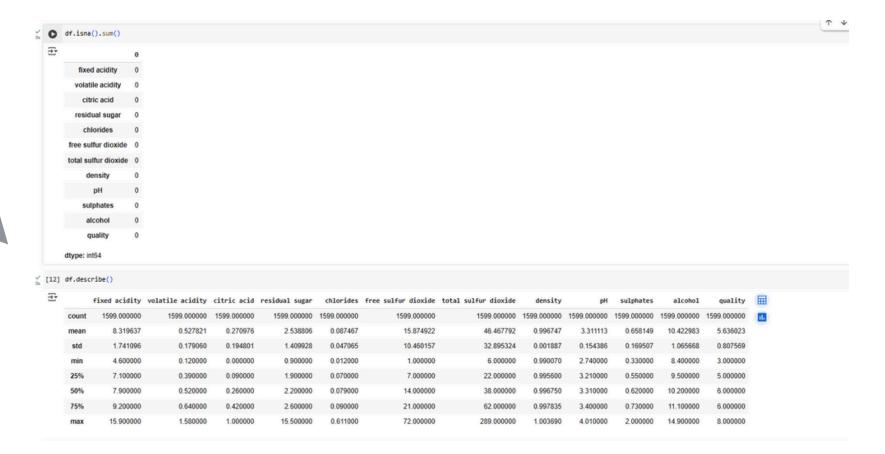
int64

1599 non-null

quality

memory usage: 150.0 KB

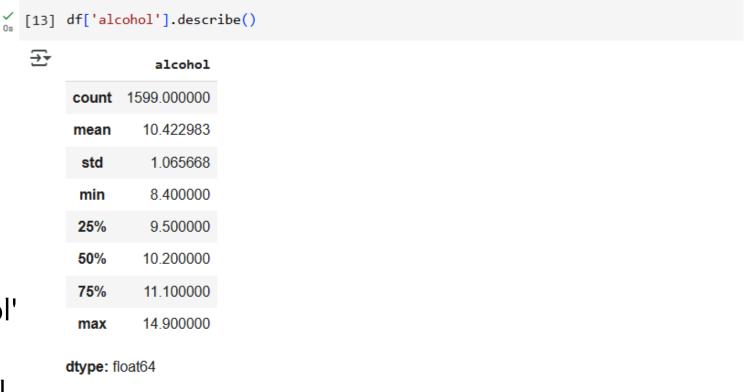
dtypes: float64(11), int64(1)



This image shows an analysis of the Wine Quality dataset using Pandas. The df.isna().sum() function confirms there are no missing values, while df.describe() provides statistical summaries such as mean, standard deviation, minimum, maximum, and percentiles to help understand the data distribution and characteristics.

The output provides a statistical summary of the 'alcohol' column in the Wine Quality dataset, indicating there are 1,599 entries with no missing values. The average alcohol content is approximately 10.42%, with a standard deviation of 1.07. The values range from a minimum of 8.4% to a maximum of 14.9%. The 25th percentile is 9.5%, the median (50th percentile) is 10.2%, and the 75th percentile is 11.1%, showing a relatively even distribution of alcohol content across the samples.

40



```
# Mengatasi missing value

for column in df.columns:

if df[column].dtype == 'object':

# Jika kolom bertipe object, isi dengan mode

df[column].fillna(df[column].mode()[0], inplace=True)

else:

# Jika kolom bertipe numerik, isi dengan mean

df[column].fillna(df[column].mean(), inplace=True)

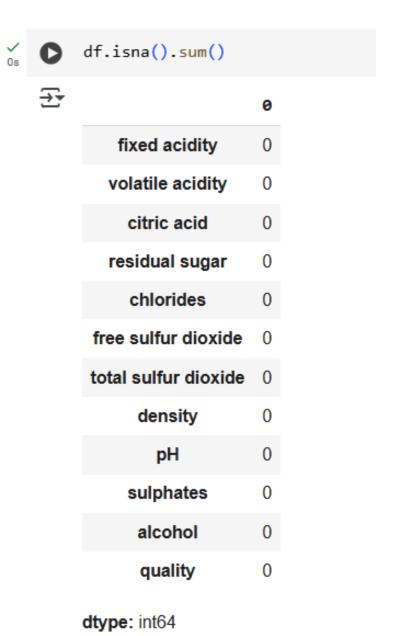
**Cipython-input-14-da1a6285f769>:8: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.

The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[coll.method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original df[column].fillna(df[column].mean(), inplace=True)
```

The code handles missing values in a DataFrame by checking each column's data type. If a column is of type object (usually for categorical/text data), it fills missing values with the most frequent value (mode). If the column is numeric, it fills missing values with the average (mean). However, a warning is shown because the code uses chained assignment with the inplace=True argument, which is discouraged in future versions of pandas (3.0 and above). Instead, it's recommended to reassign the result directly to the column using df[column] = df[column].fillna(...) for more reliable behavior.

The code df.isna().sum() is used to check for missing values in each column of the dataset. The output shows that all columns contain o missing values, indicating that the dataset is complete and does not require any imputation or cleaning for missing data. This ensures that further analysis can be performed without handling null values.



The df.info() output provides a summary of the dataset, indicating that it contains 1,599 entries and 12 columns. Each column has 1,599 non-null values, meaning there are no missing values in the dataset. The data types are mostly float64, except for the quality column which is int64. The total memory usage of the DataFrame is approximately 150.0 KB. This confirms the dataset is complete and well-structured for analysis.

```
[16] df.info()
     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 1599 entries, 0 to 1598
     Data columns (total 12 columns):
          Column
                                 Non-Null Count Dtype
          fixed acidity
                                1599 non-null
                                                 float64
          volatile acidity
                                                 float64
                                1599 non-null
          citric acid
                                1599 non-null
                                                 float64
          residual sugar
                                1599 non-null
                                                 float64
          chlorides
                                1599 non-null
                                                 float64
          free sulfur dioxide
                                                 float64
                                1599 non-null
          total sulfur dioxide 1599 non-null
                                                 float64
          density
                                1599 non-null
                                                 float64
                                1599 non-null
                                                 float64
                                                 float64
          sulphates
                                1599 non-null
          alcohol
                                1599 non-null
                                                 float64
          quality
                                1599 non-null
                                                 int64
     dtypes: float64(11), int64(1)
```

memory usage: 150.0 KB

The code checks for duplicate rows across all columns using df.duplicated().sum(). The output indicates that there are 240 duplicate entries in the dataset. This suggests that multiple rows contain identical values across all features and may need to be reviewed or removed for accurate analysis.

```
[18] # Handling duplicate

df = df.drop_duplicates()
```

The code df = df.drop_duplicates() is used to remove all duplicate rows from the dataset. By assigning the result back to df, the dataset is updated to include only unique rows, which helps improve data quality and avoid redundancy in analysis.

The code checks for any remaining duplicate rows in the dataset after handling them. Since the result shows Jumlah data yang duplikat = o, it confirms that all duplicate rows have been successfully removed and the dataset now only contains unique records.



THANKYOU

DATA SCIENCE