ibimbing

EXPLORATORY
Data Analysis

BY : UUN TRI JAYANTI

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('winequality-red.csv')
```

The code imports libraries for data analysis and visualization, then reads the `winequality-red.csv` file into the `df` variable as a DataFrame for analysis.

check if there are any blank values in any column. If there are, we fill them with the median because it is more resistant to outliers than the mean.

```
print(df.isnull().sum())
df.fillna(df.median(), inplace=True)
```

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
```

```
print("Jumlah duplikat:", df.duplicated().sum())
df.drop_duplicates(inplace=True)
```

```
Jumlah duplikat: 240
```

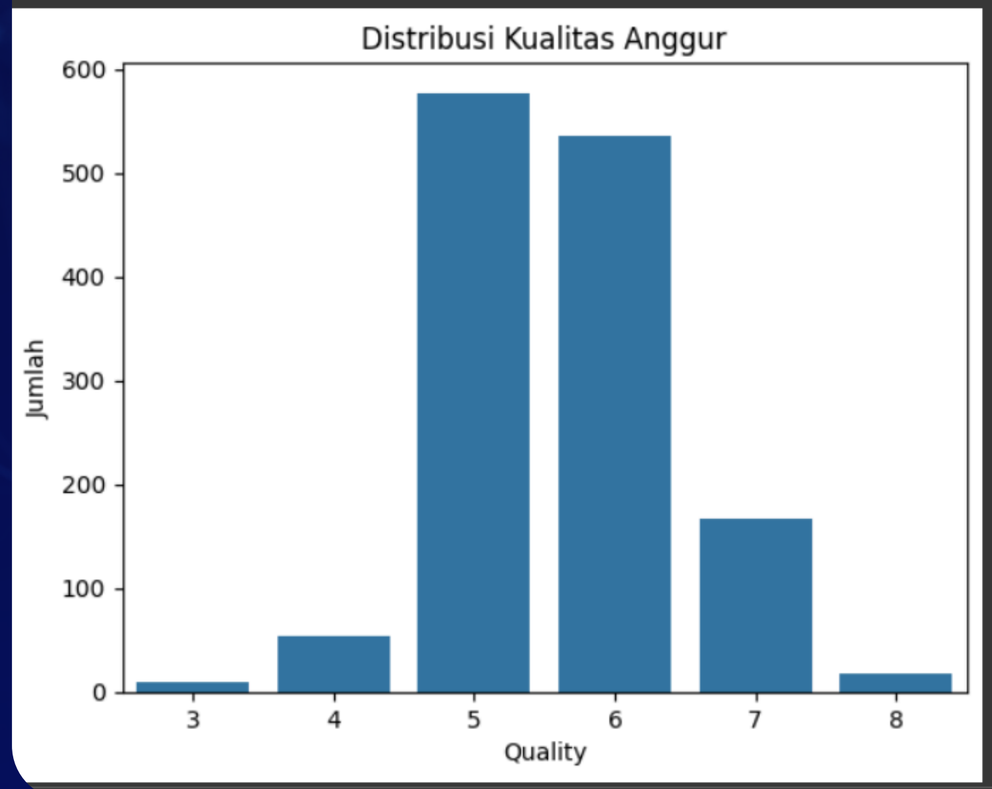
check if there are any identical rows. If found, we delete them so as not to interfere with the analysis and prediction model. This is very important to maintain data quality.

Provides a summary of the data type, number of non-null values, and statistics such as mean, standard deviation, and maximum/minimum values.

```
print(df.info())
```

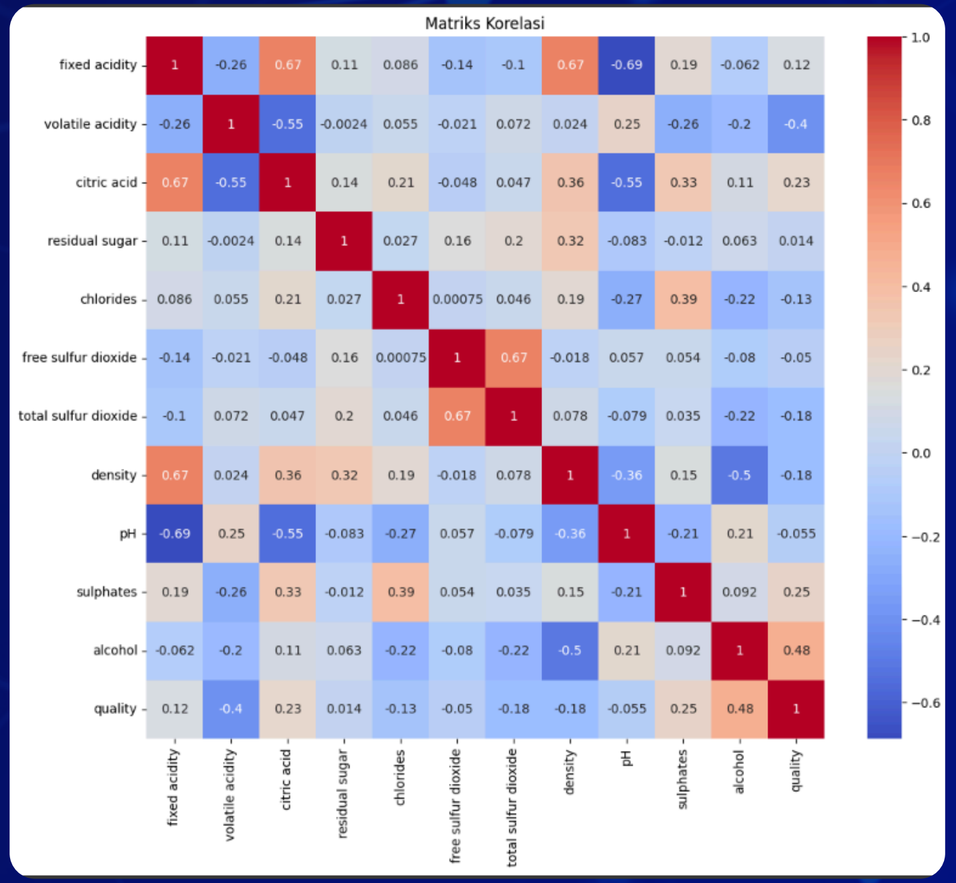
```
<class 'pandas.core.frame.DataFrame'>
Index: 1102 entries, 0 to 1597
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1102 non-null   float64
 1   volatile acidity      1102 non-null   float64
 2   citric acid           1102 non-null   float64
 3   residual sugar        1102 non-null   float64
 4   chlorides             1102 non-null   float64
 5   free sulfur dioxide   1102 non-null   float64
 6   total sulfur dioxide  1102 non-null   float64
 7   density               1102 non-null   float64
 8   pH                    1102 non-null   float64
 9   sulphates             1102 non-null   float64
 10  alcohol               1102 non-null   float64
 11  quality               1102 non-null   int64  
 12  quality_label         1102 non-null   object 
 13  quality_label_encoded 1102 non-null   int64  
dtypes: float64(11), int64(2), object(1)
memory usage: 129.1+ KB
None
```

```
sns.countplot(x='quality', data=df)
plt.title('Distribusi Kualitas Anggur')
plt.xlabel('Quality')
plt.ylabel('Jumlah')
plt.show()
```



This visualization shows how many wines fall into each quality level, helping us understand whether the target data is balanced or not.

```
plt.figure(figsize=(12,10))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title('Matriks Korelasi')
plt.show()
```



Heatmaps help us understand the relationship between variables. For example, alcohol may have a positive correlation with quality.

ibimbing

THANK YOU

-