PRACTICE 04 : STEP 2. DATA PREPROCESSING

1. Aim of the practice

- Data cleaning
- Data transformation

2. Data cleaning

2.1. Dealing with missing data

Check for missing data:

- The info() method provides a concise summary of the DataFrame, including the count of non-null values.
- df.isnull().sum()
- Using missingno package: import missingno as msno
  msno.matrix(df)

Exercise: add some missing data to cp variable by: df.loc[1, 'cp']=np.nan

Check the missing data by the previous methods.

Methods to deal with Missing Data

- df_filled = df.fillna(0)
- df_filled_mean = df.fillna(df.mean())
- df.dropna()

2.2. Dealing with outliers

Check for outliers:

1. Using z score: from scipy.stats import zscore

```
z_scores = np.abs(zscore(df))
print("Z-scores:")
print(z_scores) # Set a threshold for outliers (Z-score > 3)
outliers = (z_scores > 3)
print("\n Outliers based on Z-score:")
print(outliers)
np.sum(outliers)
```

2. Using IQR:

```
Q1 = df.quantile(0.25)
```

```
Q3 = df.quantile(0.75)
# Calculate IQR
IQR = Q3 - Q1
# Define outliers (values outside the range of Q1 - 1.5*IQR to Q3 + 1.5*IQR)
 outliers_iqr = (df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))
print(outliers_iqr)
```

Methods to deal with outliers

```
df_replaced = df.copy()
for column in df.columns:
df_replaced[column] = np.where(outliers_iqr[column], df[column].median(), df[column])
  print("\nDataFrame after replacing outliers with median values:")
print(df_replaced)
```

3. Data transformation

Common transformations include normalizing, scaling, encoding categorical variables, and applying mathematical transformations.

Normalization (Min-Max Scaling): Rescales data to a [0, 1] range.

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
# Initialize MinMaxScaler
scaler = MinMaxScaler()
# Apply MinMax scaling
df_normalized = pd.DataFrame(scaler.fit_transform(df), columns=df.columns)
print("Data after Min-Max Normalization:")
print(df_normalized)
```