import csv

import pandas as pd

# Step 1: Create a Sample Dataset with Anomalies

data = [

["id", "name", "age", "salary"],

[1, "Alice", 25, 50000],

[2, "Bob", 30, None], # Missing salary

[3, "Charlie", 35, 70000],

[4, "David", None, 60000], # Missing age

[5, "Eve", 22, 55000],

[6, "Frank", 40, 80000],

[7, "Grace", 30, 65000],

[8, "Hank", 25, 300000], # Outlier salary

[9, "Ivy", 28, 50000],

[10, "John", 35, 70000],

[2, "Bob", 30, None], # Duplicate entry

]

with open('sample\_dataset.csv', mode='w', newline='') as file:

writer = csv.writer(file)

writer.writerows(data)

# Step 2: Load the Dataset

data = []

with open('sample\_dataset.csv', mode='r') as file:

reader = csv.reader(file)

for row in reader:

data.append(row)

columns = data[0]

rows = data[1:]

df = pd.DataFrame(rows, columns=columns)

# Step 3: Clean the Anomalies

# Convert columns to appropriate data types

df['id'] = pd.to\_numeric(df['id'])

df['age'] = pd.to\_numeric(df['age'])

df['salary'] = pd.to\_numeric(df['salary'])

# Remove duplicates

df = df.drop\_duplicates()

# Handle missing values

df['age'] = df['age'].fillna(df['age'].mean())

df['salary'] = df['salary'].fillna(df['salary'].median())

# Handle outliers in 'salary' using the IQR method

Q1 = df['salary'].quantile(0.25)

Q3 = df['salary'].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

# Replace outliers with the median salary

median\_salary = df['salary'].median()

df.loc[(df['salary'] < lower\_bound) | (df['salary'] > upper\_bound), 'salary'] = median\_salary

# Display the cleaned DataFrame

print(df)