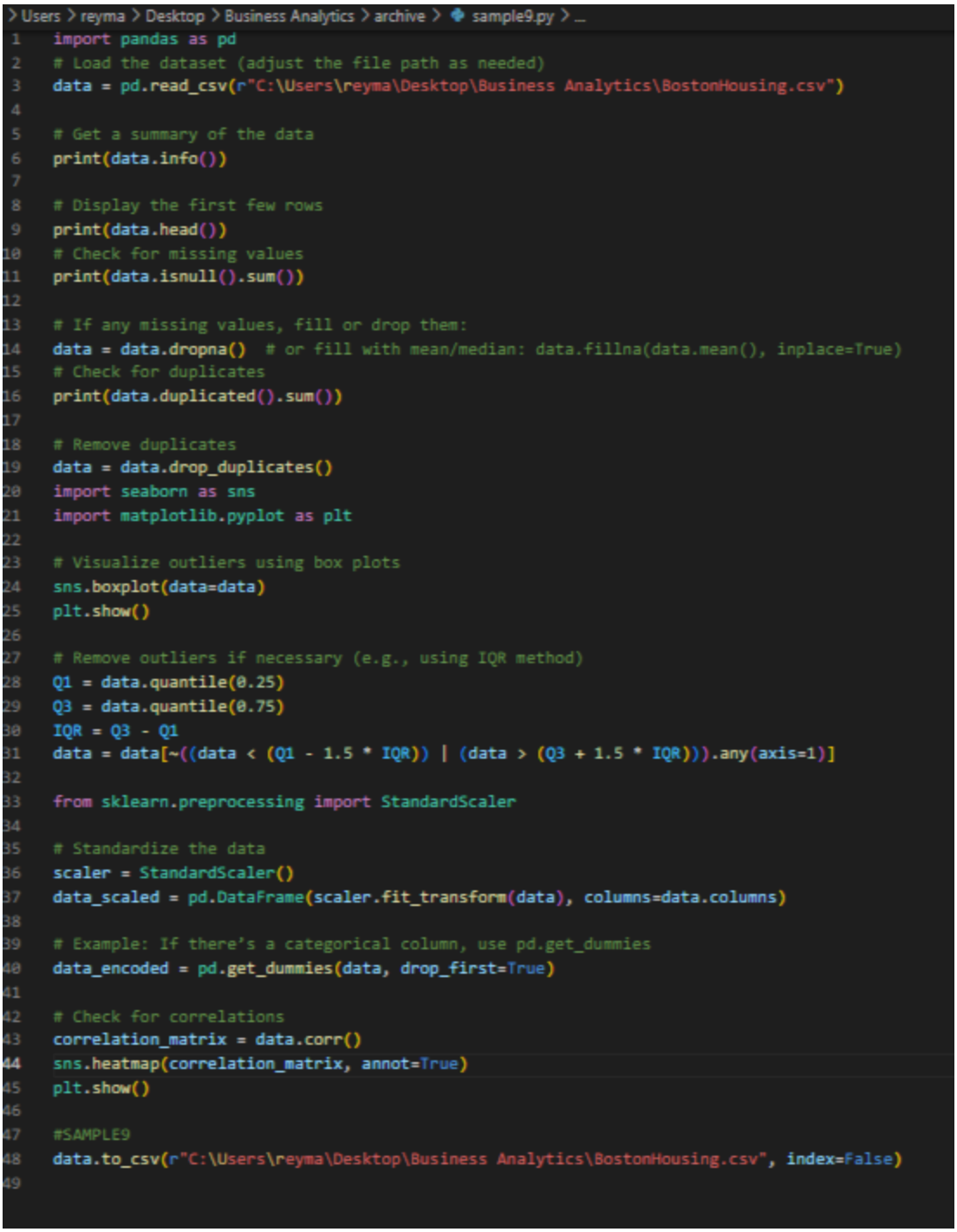

Code:

import pandas as pd

Load the dataset (adjust the file path as needed)

data = pd.read_csv(r"C:\Users\reyma\Desktop\Business Analytics\BostonHousing.csv")

```
# Get a summary of the data
print(data.info())

# Display the first few rows
print(data.head())
# Check for missing values
print(data.isnull().sum())

# If any missing values, fill or drop them:
data = data.dropna()  # or fill with mean/median: data.fillna(data.mean(), inplace=True)
# Check for duplicates
print(data.duplicated().sum())

# Remove duplicates
data = data.drop_duplicates()
import seaborn as sns
import matplotlib.pyplot as plt

# Visualize outliers using box plots
sns.boxplot(data=data)
plt.show()

# Remove outliers if necessary (e.g., using IQR method)
Q1 = data.quantile(0.25)
Q3 = data.quantile(0.75)
IQR = Q3 - Q1
data = data[~((data < (Q1 - 1.5 * IQR)) | (data > (Q3 + 1.5 * IQR))).any(axis=1)]
```

```
from sklearn.preprocessing import StandardScaler

# Standardize the data
scaler = StandardScaler()
data_scaled = pd.DataFrame(scaler.fit_transform(data), columns=data.columns)

# Example: If there’s a categorical column, use pd.get_dummies
data_encoded = pd.get_dummies(data, drop_first=True)

# Check for correlations
correlation_matrix = data.corr()
sns.heatmap(correlation_matrix, annot=True)
plt.show()

#SAMPLE9
data.to_csv(r"C:\Users\reyma\Desktop\Business Analytics\BostonHousing.csv", index=False)
```