

```

> Users > reyma > Desktop > Business Analytics > archive > sample9.py > ...
1  import pandas as pd
2  # Load the dataset (adjust the file path as needed)
3  data = pd.read_csv(r"C:\Users\reyma\Desktop\Business Analytics\BostonHousing.csv")
4
5  # Get a summary of the data
6  print(data.info())
7
8  # Display the first few rows
9  print(data.head())
10 # Check for missing values
11 print(data.isnull().sum())
12
13 # If any missing values, fill or drop them:
14 data = data.dropna() # or fill with mean/median: data.fillna(data.mean(), inplace=True)
15 # Check for duplicates
16 print(data.duplicated().sum())
17
18 # Remove duplicates
19 data = data.drop_duplicates()
20 import seaborn as sns
21 import matplotlib.pyplot as plt
22
23 # Visualize outliers using box plots
24 sns.boxplot(data=data)
25 plt.show()
26
27 # Remove outliers if necessary (e.g., using IQR method)
28 Q1 = data.quantile(0.25)
29 Q3 = data.quantile(0.75)
30 IQR = Q3 - Q1
31 data = data[~((data < (Q1 - 1.5 * IQR)) | (data > (Q3 + 1.5 * IQR))).any(axis=1)]
32
33 from sklearn.preprocessing import StandardScaler
34
35 # Standardize the data
36 scaler = StandardScaler()
37 data_scaled = pd.DataFrame(scaler.fit_transform(data), columns=data.columns)
38
39 # Example: If there's a categorical column, use pd.get_dummies
40 data_encoded = pd.get_dummies(data, drop_first=True)
41
42 # Check for correlations
43 correlation_matrix = data.corr()
44 sns.heatmap(correlation_matrix, annot=True)
45 plt.show()
46
47 #SAMPLE9
48 data.to_csv(r"C:\Users\reyma\Desktop\Business Analytics\BostonHousing.csv", index=False)
49

```

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)
```