NAVODAYA INSTITUTE OF TECHNOLOGY

MACHINE LEARNING LAB (BCSL606)

Program 1

1. Develop a program to create histograms for all numerical features and analyze the distribution of each feature. Generate box plots for all numerical features and identify any outliers. Use California Housing dataset.

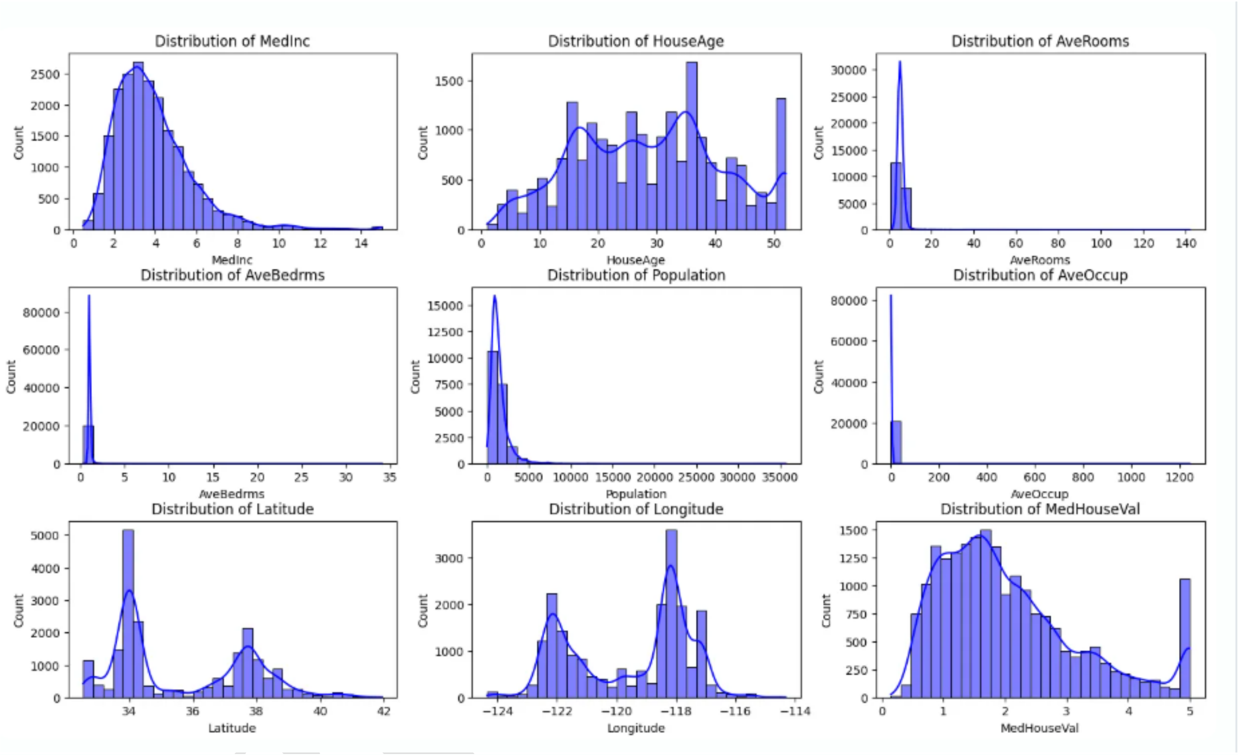
PROGRAM:

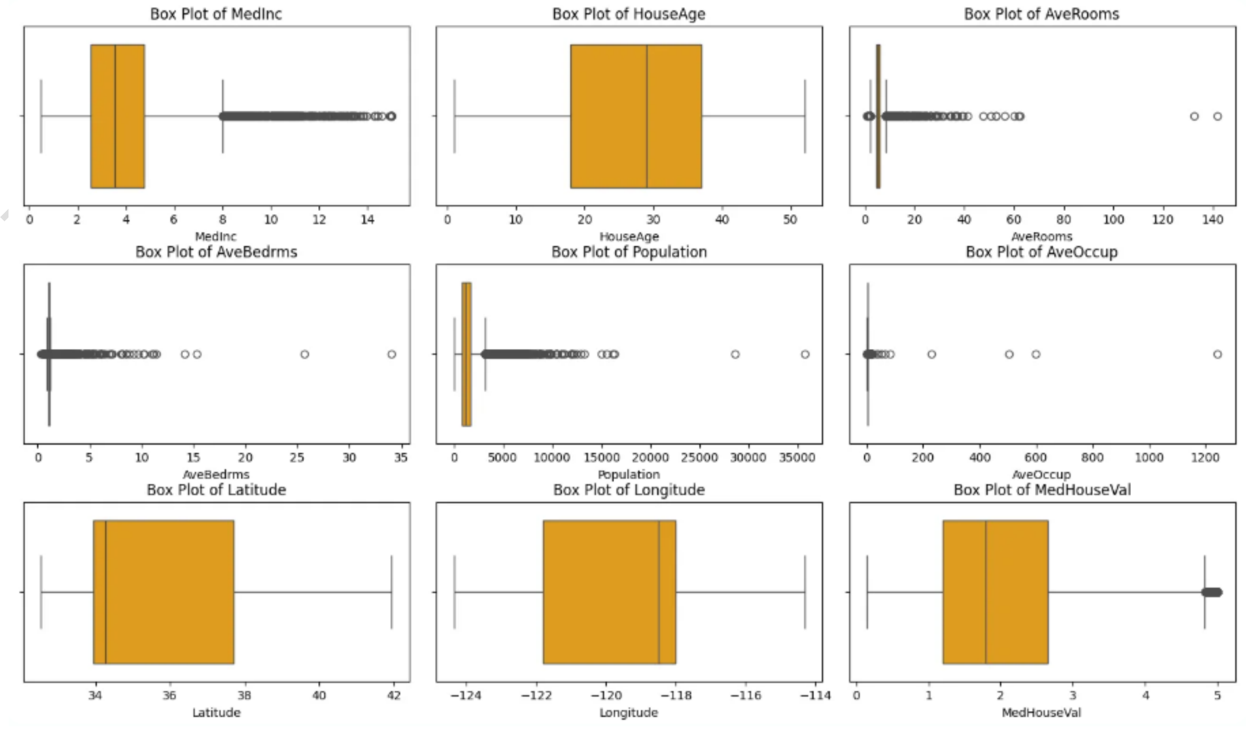
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.datasets import fetch_california_housing



# Step 1: Load the California Housing dataset

data = fetch_california_housing(as_frame=True)

housing_df = data.frame



# Step 2: Create histograms for numerical features
```

```
numerical_features = housing_df.select_dtypes(include=[np.number]).columns


# Plot histograms
plt.figure(figsize=(15, 10))
for i, feature in enumerate(numerical_features):
    plt.subplot(3, 3, i + 1)
    sns.histplot(housing_df[feature], kde=True, bins=30, color='blue')
    plt.title(f'Distribution of {feature}')
plt.tight_layout()
plt.show()


# Step 3: Generate box plots for numerical features
plt.figure(figsize=(15, 10))
for i, feature in enumerate(numerical_features):
    plt.subplot(3, 3, i + 1)
    sns.boxplot(x=housing_df[feature], color='orange')
    plt.title(f'Box Plot of {feature}')
plt.tight_layout()
plt.show()
```

```
# Step 4: Identify outliers using the IQR method
print("Outliers Detection:")
outliers_summary = {}
for feature in numerical_features:
    Q1 = housing_df[feature].quantile(0.25)
    Q3 = housing_df[feature].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = housing_df[(housing_df[feature] < lower_bound) | (housing_df[feature] >
upper_bound)]
    outliers_summary[feature] = len(outliers)
    print(f"{feature}: {len(outliers)} outliers")


# Optional: Print a summary of the dataset
print("\nDataset Summary:")
print(housing_df.describe())
```

OUTPUT:


Distribution of MedInc
Count
MedInc
0 2 4 6 8 10 12 14
0 500 1000 1500 2000 2500
Distribution of HouseAge
Count
HouseAge
0 10 20 30 40 50
0 500 1000 1500
Distribution of AveRooms
Count
AveRooms
0 20 40 60 80 100 120 140
0 5000 10000 15000 20000 25000 30000
Distribution of AveBedrms
Count
AveBedrms
0 5 10 15 20 25 30 35
0 20000 40000 60000 80000
Distribution of Population
Count
Population
0 5000 10000 15000 20000 25000 30000 35000
0 2500 5000 7500 10000 12500 15000
Distribution of AveOccup
Count
AveOccup
0 200 400 600 800 1000 1200
0 20000 40000 60000 80000
Distribution of Latitude
Count
Latitude
34 36 38 40 42
0 1000 2000 3000 4000 5000
Distribution of Longitude
Count
Longitude
−124 −122 −120 −118 −116 −114
0 1000 2000 3000
Distribution of MedHouseVal
Count
MedHouseVal
0 1 2 3 4 5
0 250 500 750 1000 1250 1500


Box Plot of MedInc
MedInc
0 2 4 6 8 10 12 14
Box Plot of HouseAge
HouseAge
0 10 20 30 40 50
Box Plot of AveRooms
AveRooms
0 20 40 60 80 100 120 140
Box Plot of AveBedrms
AveBedrms
0 5 10 15 20 25 30 35
Box Plot of Population
Population
0 5000 10000 15000 20000 25000 30000 35000
Box Plot of AveOccup
AveOccup
0 200 400 600 800 1000 1200
Box Plot of Latitude
Latitude
34 36 38 40 42
Box Plot of Longitude
Longitude
−124 −122 −120 −118 −116 −114
Box Plot of MedHouseVal
MedHouseVal
0 1 2 3 4 5

Outliers Detection:

MedInc: 681 outliers

HouseAge: 0 outliers

AveRooms: 511 outliers

AveBedrms: 1424 outliers

Population: 1196 outliers

AveOccup: 711 outliers

Latitude: 0 outliers

Longitude: 0 outliers

MedHouseVal: 1071 outliers

Dataset Summary:

```
          MedInc      HouseAge  ...     Longitude   MedHouseVal
count  20640.000000  20640.000000  ...  20640.000000  20640.000000
mean       3.870671     28.639486  ...   -119.569704      2.068558
std        1.899822     12.585558  ...      2.003532      1.153956
min        0.499900      1.000000  ...   -124.350000      0.149990
25%        2.563400     18.000000  ...   -121.800000      1.196000
50%        3.534800     29.000000  ...   -118.490000      1.797000
75%        4.743250     37.000000  ...   -118.010000      2.647250
```

```
max       15.000100     52.000000  ...   -114.310000       5.000010
```