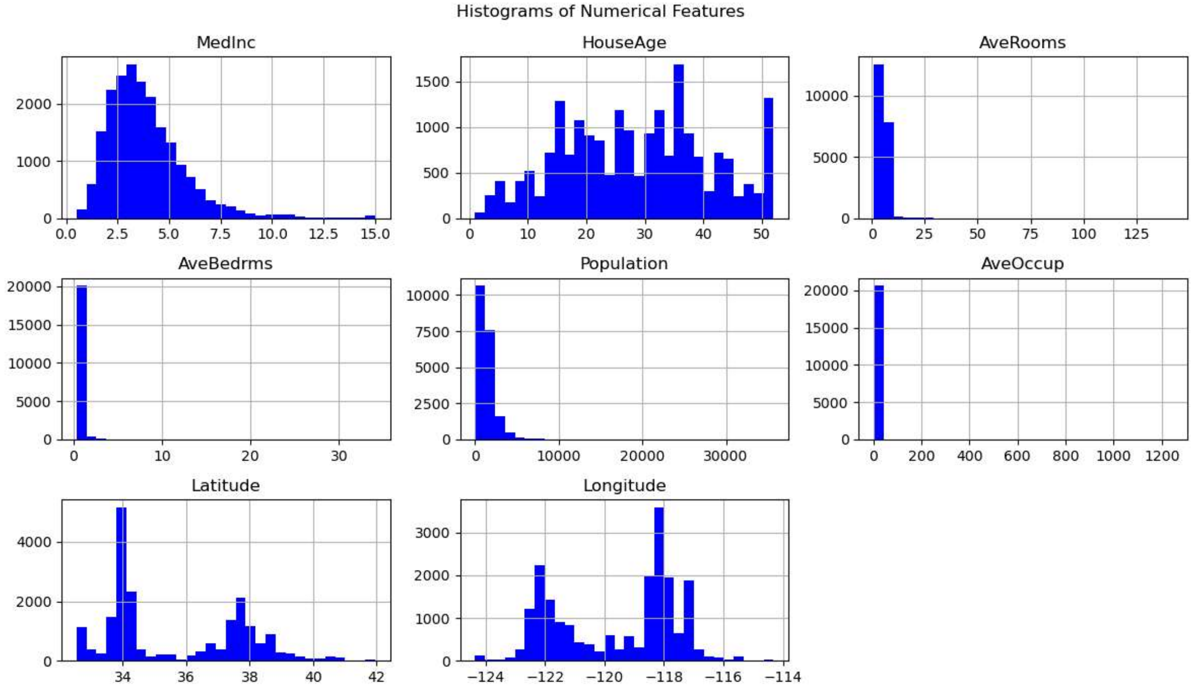
EXPERIMENT : 1

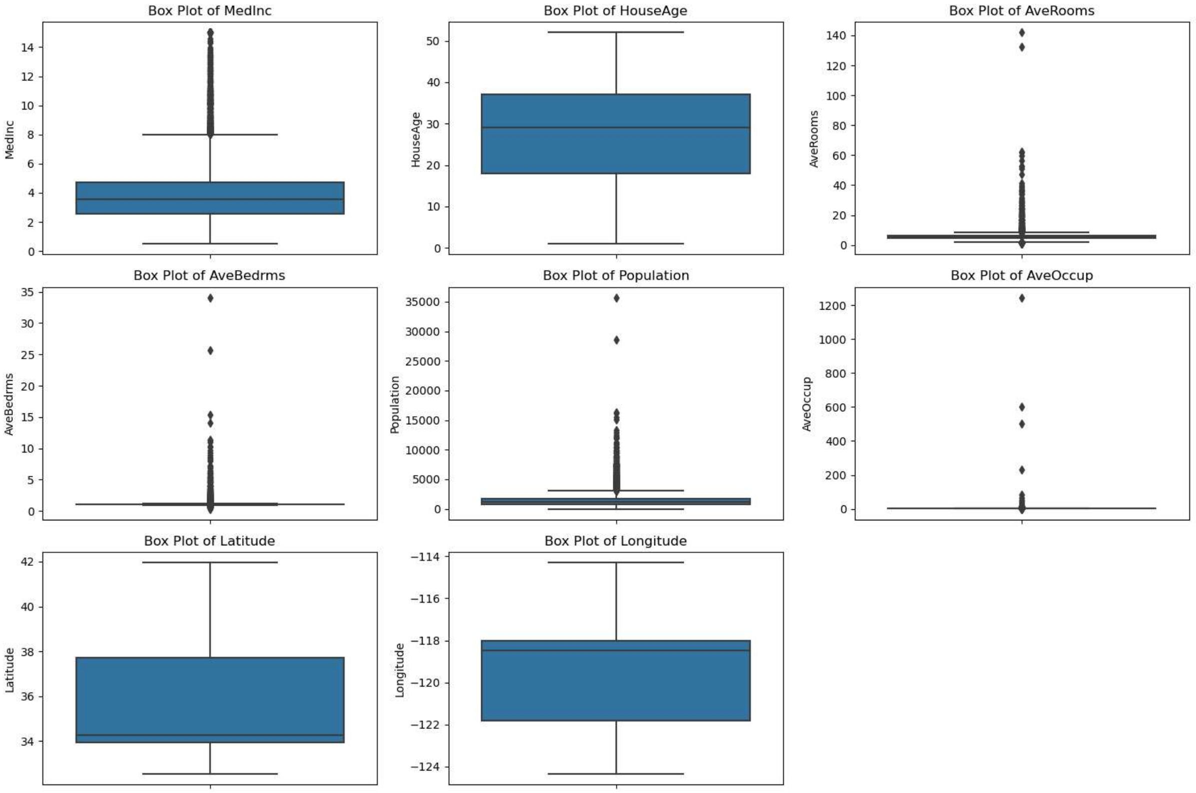
Develop a program to create histograms for all numerical features and analyze the :distribution of each feature. Generate box plots for all numerical features and identify any outliers. Use California Housing dataset.

In [1]:

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np
from sklearn.datasets import fetch_california_housing

# Load the California Housing dataset
california_housing = fetch_california_housing()
data = pd.DataFrame(california_housing.data, columns=california_housing.feature_names)

# checking for numerical features
numerical_features = data.select_dtypes(include=[np.number]).columns
print(numerical_features)

# Create histograms for all numerical features
data.hist(bins=30, figsize=(12, 7),color='blue')
plt.suptitle('Histograms of Numerical Features')
plt.tight_layout()
plt.show()

# Create box plots for all numerical features
plt.figure(figsize=(15, 10))
for i, column in enumerate(data.columns, 1):
    plt.subplot(3, 3, i)
    sns.boxplot(y=data[column])
    plt.title(f'Box Plot of {column}')
plt.tight_layout()
plt.show()

# Obtain and print outliers summary
print("Outliers Detection:\n")
outliers_summary = {}
for feature in numerical_features:
    Q1 = data[feature].quantile(0.25)
    Q3 = data[feature].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = data[(data[feature] < lower_bound) | (data[feature] > upper_bound)]
    outliers_summary[feature] = len(outliers)
    print(f"\t{feature}: {len(outliers)} outliers\t")
```

```
Index(['MedInc', 'HouseAge', 'AveRooms', 'AveBedrms', 'Population', 'AveOccup',
       'Latitude', 'Longitude'],
      dtype='object')
```





```
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
```

```
In [ ]:
```