In [2]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

```

In [3]:

```python
df = pd.read_csv("nces330_20.csv")
df
```

```
      Year    State                Type  Length      Expense  Value
   0  2013  Alabama             Private  4-year  Fees/Tuition  13983
   1  2013  Alabama             Private  4-year    Room/Board   8503
   2  2013  Alabama     Public In-State  2-year  Fees/Tuition   4048
   3  2013  Alabama     Public In-State  4-year  Fees/Tuition   8073
   4  2013  Alabama     Public In-State  4-year    Room/Board   8473
 ...   ...      ...                 ...     ...           ...    ...
3543  2021  Wyoming     Public In-State  2-year  Fees/Tuition   3987
3544  2021  Wyoming     Public In-State  4-year    Room/Board   9799
3545  2021  Wyoming Public Out-of-State  2-year  Fees/Tuition   9820
3546  2021  Wyoming Public Out-of-State  4-year  Fees/Tuition  14710
3547  2021  Wyoming Public Out-of-State  4-year    Room/Board   9799
```

In [7]:

```
#lets look at the most recent data (year 2021) and fees/tuition only
df_yr2021_4Year_tuition = df[(df['Year'] == 2021) & (df['Expense'] == 'F
                                (df['Length'] == '4-year')]
df_yr2021_4Year_tuition
```

Out[7]:

	Year	State	Type	Length	Expense	Value
3203	2021	Alabama	Private	4-year	Fees/Tuition	17354
3208	2021	Alabama	Public Out-of-State	4-year	Fees/Tuition	27005
3210	2021	Alaska	Private	4-year	Fees/Tuition	19575
3213	2021	Alaska	Public Out-of-State	4-year	Fees/Tuition	25535
3215	2021	Arizona	Private	4-year	Fees/Tuition	13108
...	...	...	...	...	...	...
3529	2021	West Virginia	Private	4-year	Fees/Tuition	12413
3534	2021	West Virginia	Public Out-of-State	4-year	Fees/Tuition	22475
3536	2021	Wisconsin	Private	4-year	Fees/Tuition	35674
3541	2021	Wisconsin	Public Out-of-State	4-year	Fees/Tuition	26970
3546	2021	Wyoming	Public Out-of-State	4-year	Fees/Tuition	14710

101 rows × 6 columns

In [8]:

```python
df_yr2019_4Year_tuition = df[(df['Year'] == 2019) & (df['Expense'] == 'F
                                    (df['Length'] == '4-year')]
df_yr2019_4Year_tuition
```

Out[8]:

	Year	State	Type	Length	Expense	Value
2411	2019	Alabama	Private	4-year	Fees/Tuition	16119
2414	2019	Alabama	Public In-State	4-year	Fees/Tuition	10138
2417	2019	Alabama	Public Out-of-State	4-year	Fees/Tuition	25782
2419	2019	Alaska	Private	4-year	Fees/Tuition	19315
2421	2019	Alaska	Public In-State	4-year	Fees/Tuition	8396
...	...	...	...	...	...	...
2793	2019	Wisconsin	Private	4-year	Fees/Tuition	34424
2796	2019	Wisconsin	Public In-State	4-year	Fees/Tuition	8697
2799	2019	Wisconsin	Public Out-of-State	4-year	Fees/Tuition	25063
2802	2019	Wyoming	Public In-State	4-year	Fees/Tuition	4596

In [9]:

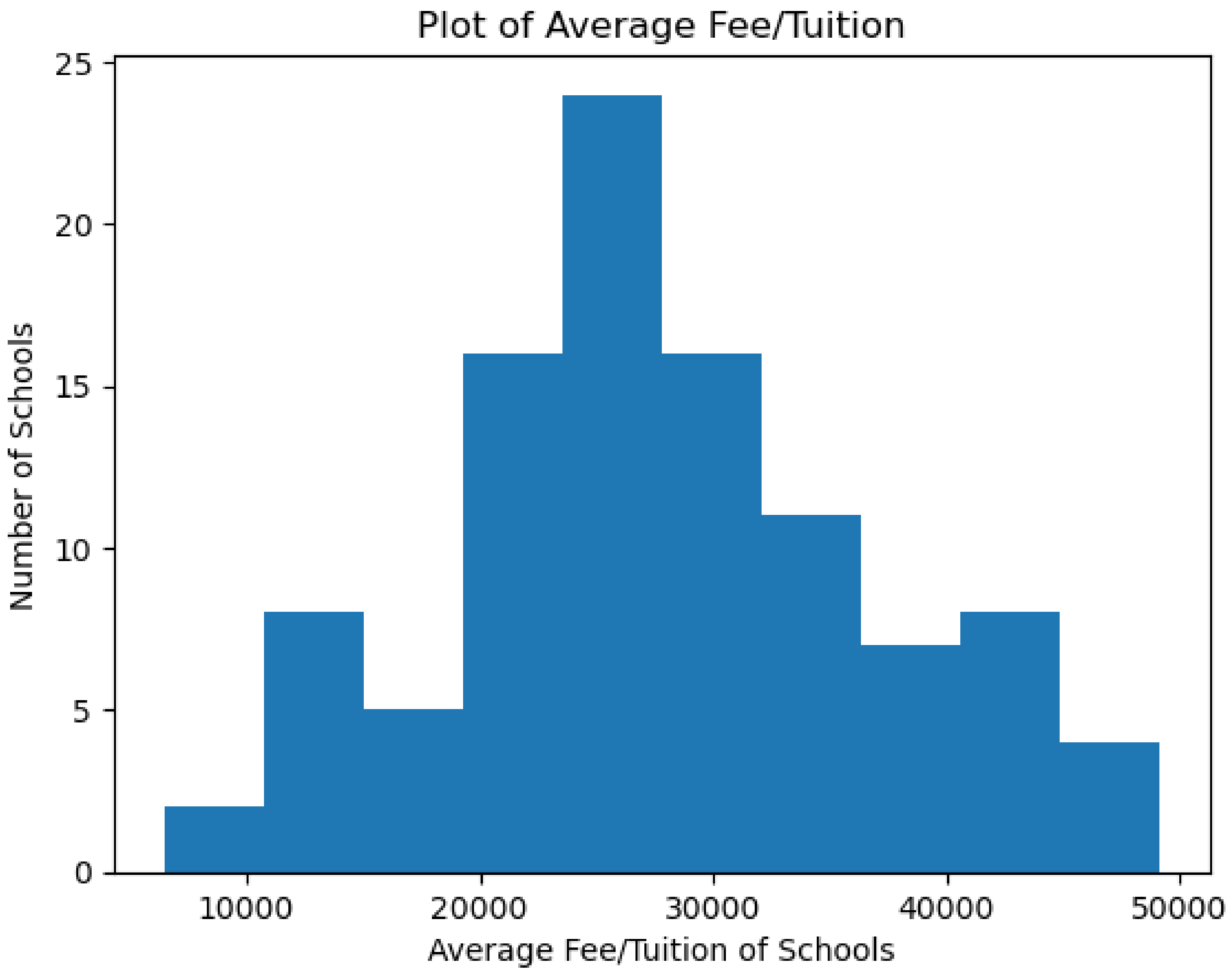
```python
df_yr2019_4Year_tuition['Type'].unique()
```

Out[9]:

```
array(['Private', 'Public In-State', 'Public Out-of-State'], d
type=object)
```

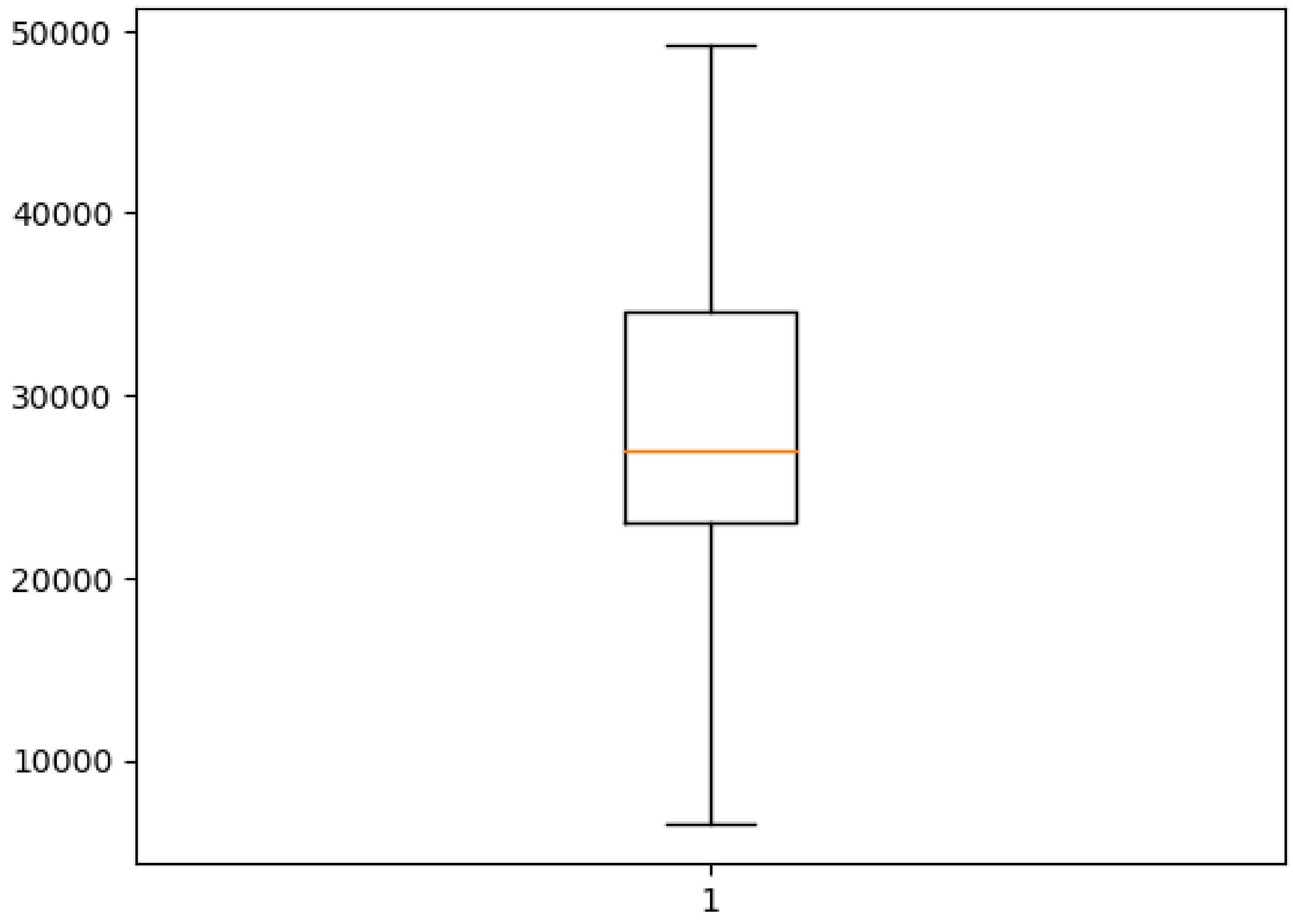
In [10]:

```python
plt.hist(df_yr2021_4Year_tuition['Value'])
plt.title("Plot of Average Fee/Tuition")
plt.xlabel("Average Fee/Tuition of Schools")
plt.ylabel("Number of Schools")
plt.show()
```



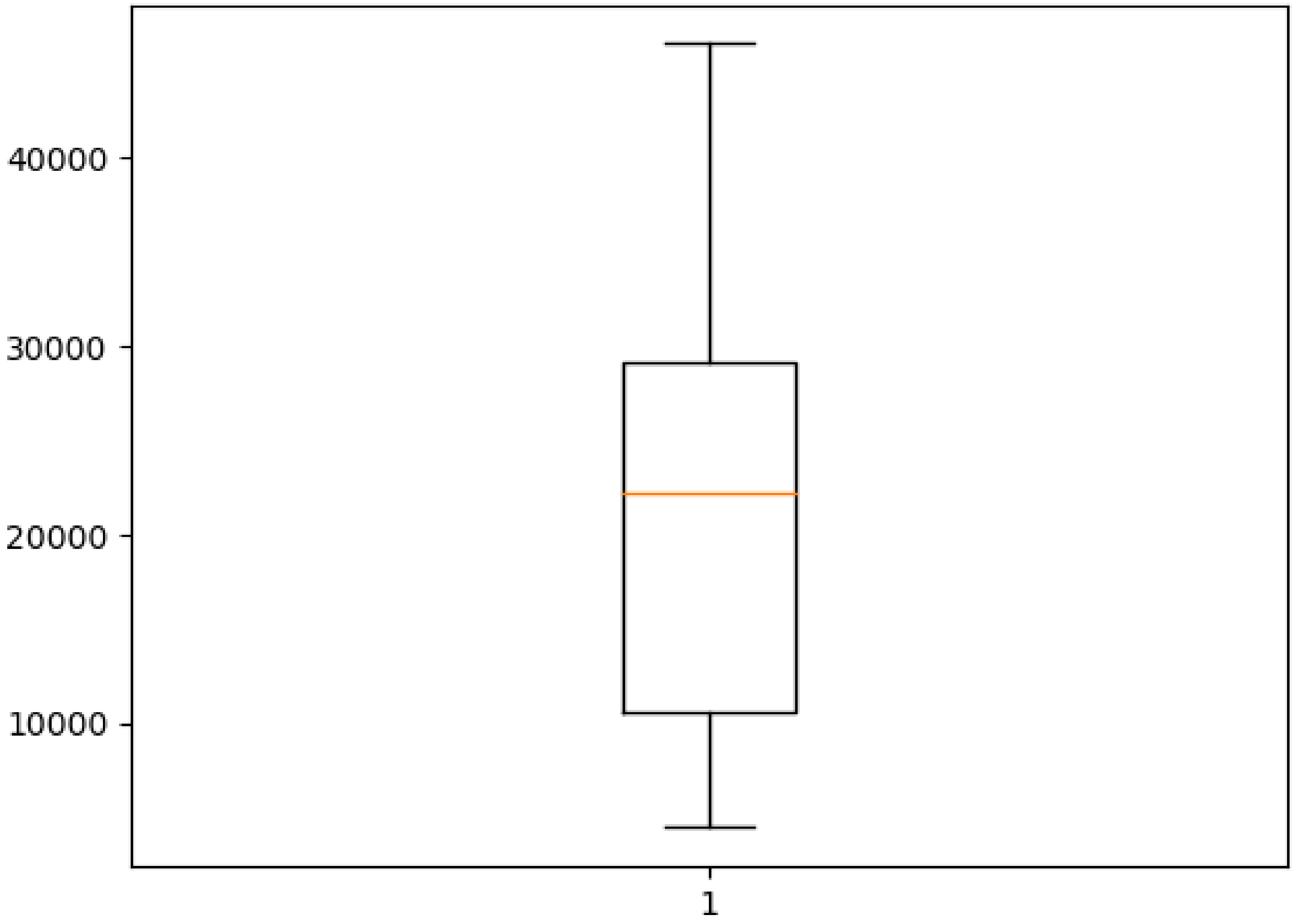
In [11]:

```
plt.boxplot(df_yr2021_4Year_tuition['Value'])
plt.show()
```



In [14]:

```
plt.boxplot(df_yr2019_4Year_tuition['Value'])
plt.show()
```



In [15]:

```
df_yr2021_avgTuit = df_yr2021_4Year_tuition['Value'].groupby([df.iloc[:,
print("mean tuition of 4 year institutions in year 2021 is ",df_yr2021_a

```

```
mean tuition of 4 year institutions in year 2021 is  28263.722
77227723  usd
```

In [16]:

```
df_yr2019_avgTuit = df_yr2019_4Year_tuition['Value'].groupby([df.iloc[:,
print('mean tuition of 4 year institutions in year 2019 is ', df_yr2019_
```

```
mean tuition of 4 year institutions in year 2019 is  21300.453
94736842  usd
```

Statistical Study: We have two means in this study. The mean of the average tution costs of 2019 and the mean of the average tuition cost of 2021.

Lets obtain a few more pieces of information necessary for completin

In [17]:

```python
df_yr2021_stdTuit = df_yr2021_4Year_tuition['Value'].std()
df_yr2019_stdTuit = df_yr2019_4Year_tuition['Value'].std()

df_yr2021_sampleSize = df_yr2021_4Year_tuition['Value'].count()
df_yr2019_sampleSize = df_yr2019_4Year_tuition['Value'].count()
```

In [18]:

```python
from tabulate import tabulate

info = {'Schools':['Post-Secondaries of 2021', 'Post-Secondaries of 2019
        'Sample Size' :[df_yr2021_sampleSize,df_yr2019_sampleSize],
        'Mean Tuition': [df_yr2021_avgTuit,df_yr2019_avgTuit ],
        'Sample STD':[df_yr2021_stdTuit,df_yr2019_stdTuit]}

print(tabulate(info, headers='keys', tablefmt='fancy_grid'))

```

```
╒══════════════════════════╤═══════════════╤════════════════╤
═════════════╕
│ Schools                  │   Sample Size │   Mean Tuition │
Sample STD │
╞══════════════════════════╪═══════════════╪════════════════╪
═════════════╡
│ Post-Secondaries of 2021 │           101 │        28263.7 │
9150.46 │
├──────────────────────────┼───────────────┼────────────────┼
─────────────┤
│ Post-Secondaries of 2019 │           152 │        21300.5 │
10838.6   │
╘══════════════════════════╧═══════════════╧════════════════╧
═════════════╛
```

Since we do not have a population standard deviation, the best method of testing the two means would be with the use of the t-test.

- first, we will carry out an F test to see whether we should use a pooled or non-pooled t-test

In [19]:

```
import math
import scipy.stats



print('null hypothesis : std1 = std2 \nalt hypothesis: std1 ')


```

```
null hypothesis : std1 = std2
alt hypothesis: std1
```

In [20]:

```
#the standard deviation of 2019
sa = df_yr2019_stdTuit
#the standard deviation of 2021
sb = df_yr2021_stdTuit

#F test
F = (sa)**2 / (sb)**2
round(F,3)

FcritVal = scipy.stats.f.ppf(q=1-.05, dfn=df_yr2019_sampleSize-1 , dfd=d
print('F test statistic is ', F, '\nThe F critical value is ', FcritVal)
print()
print('Since F > Fa/2, we reject null hypothesis. Therefore, we must con
```

```
F test statistic is  1.4029993700883665
The F critical value is  1.358624125232962

Since F > Fa/2, we reject null hypothesis. Therefore, we must
conduct a non-pooled t-test
```

```
Creating the hypothesis.

According to College Board articles, the USA experienced 1.8% -
8.3% increase to 4-year study tuiton, depending on the type of
school. This is for the year 2021-2022.

Here, we will hypothesize that the schools in 2021 experienced an
increase to their tuition compared to 2019. This likely being a
result of inflation and the pandemic.

H0: The tuition of schools in 2021 is the same as schools in 2019

H1: The tuition of schools in 2021 is higher than the schools in
2019.

```

```
Alpha = 0.05
```

In [36]:

```
data_group1 = df_yr2021_4Year_tuition['Value']
data_group2 = df_yr2019_4Year_tuition['Value']
data_group1
data_group2
scipy.stats.ttest_ind(data_group1, data_group2, equal_var=False, alterna
```

Out[36]:

```
Ttest_indResult(statistic=5.501718366600542, pvalue=4.86594209
77386296e-08)
```

Now to obtain the critical value. The degrees of freedom will be 100 and critical level is 0.05

In [37]:

```
scipy.stats.t.ppf(0.05, 100)
```

Out[37]:

```
-1.6602343260657506
```

To reject H0 of a right tail test, the test statistic must be bigger than critical value.

Test statistic = 5.501718366600542 Critical value = -1.6602343260657506 t.s > ta

According to this test we can reject the null hypothesis. Therefore, the tuition cost of 4-year post-secondary schools in 2021 is greater than the ones in 2019.