In this notebook, We are showing the participation counts for each state of United States

In [32]:

```
import numpy as np
from sklearn.cluster import DBSCAN
from sklearn.datasets import make_blobs
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
%matplotlib inline
```

Load the dataset Let's load the dataset containing participatng cities of US in NCAA.

In [33]:

```
import csv
import pandas as pd
import numpy as np

#Read csv
pdf = pd.read_csv("2020-Mens-Data/MDataFiles_Stage1/Cities.csv")
pdf.head(5)
```

Out[33]:

	CityID	City	State
0	4001	Abilene	TX
1	4002	Akron	OH
2	4003	Albany	NY
3	4004	Albuquerque	NM
4	4005	Allentown	PA

In [34]:

```
US_State = states = ["AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DC", "DE", "FL", "GA",
          "HI", "ID", "IL", "IN", "IA", "KS", "KY", "LA", "ME", "MD",
          "MA", "MI", "MN", "MS", "MO", "MT", "NE", "NV", "NH", "NJ",
          "NM", "NY", "NC", "ND", "OH", "OK", "OR", "PA", "RI", "SC",
          "SD", "TN", "TX", "UT", "VT", "VA", "WA", "WV", "WI", "WY"]
play_count=[]
```

participating City Count for each state

Let's count how many cities have been participated from each state?

In [35]:

```
for i in range(len(US_State)):
    #print(US_State[i])
    # row in which value of 'Age' column is more than 30
    seriesObj = pdf.apply(lambda x: True if x['State'] == US_State[i] else False , axis=1)

    # Count number of True in series
    numOfRows = len(seriesObj[seriesObj == True].index)
    play_count.append(numOfRows)

    #print('Number of Rows in dataframe in which  : ', numOfRows)
```

Visualize the count for each state

In [40]:

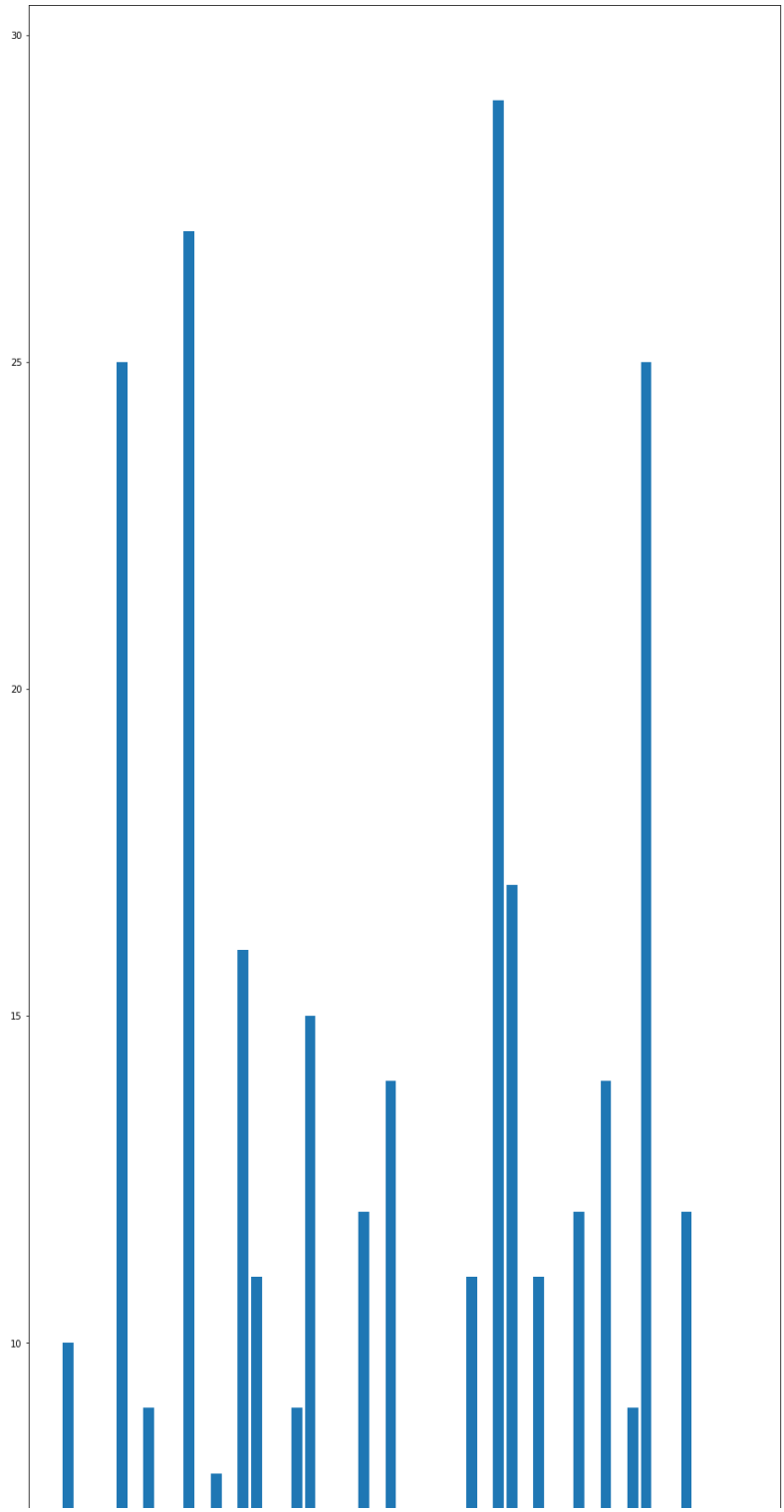
In [40]:

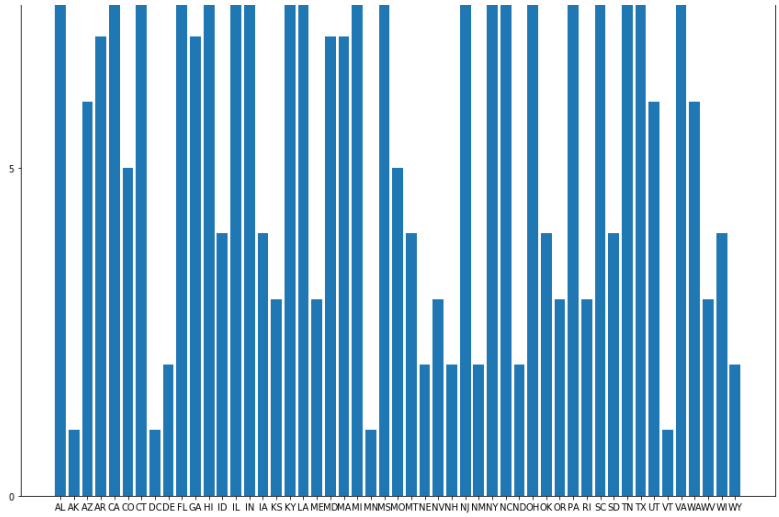
```
names = US_State
values = play_count

plt.figure(figsize=(50, 40))

plt.subplot(131)
plt.bar(names, values)
plt.suptitle('Categorical Plotting')
plt.show()
```

Categorical Plotting





Some state have very high participation

From the visualization, we can see some state has very high participation. Let's try to sort our findings and find out more insights.

In [41]:

```
print(play_count)
```

```
[10, 1, 6, 7, 25, 5, 9, 1, 2, 27, 7, 8, 4, 16, 11, 4, 3, 9, 15, 3, 7, 7, 12, 1, 14, 5, 4, 2, 3, 2,
11, 2, 29, 17, 2, 11, 4, 3, 12, 3, 14, 4, 9, 25, 6, 1, 12, 6, 3, 4, 2]
```

In [42]:

```
data_tuples = list(zip(US_State,play_count))
```

In [43]:

```
df = pd.DataFrame(data_tuples, columns=['State','City Participation Count'])
```

In [46]:

```
df = df.sort_values(by='City Participation Count', ascending=False)
```

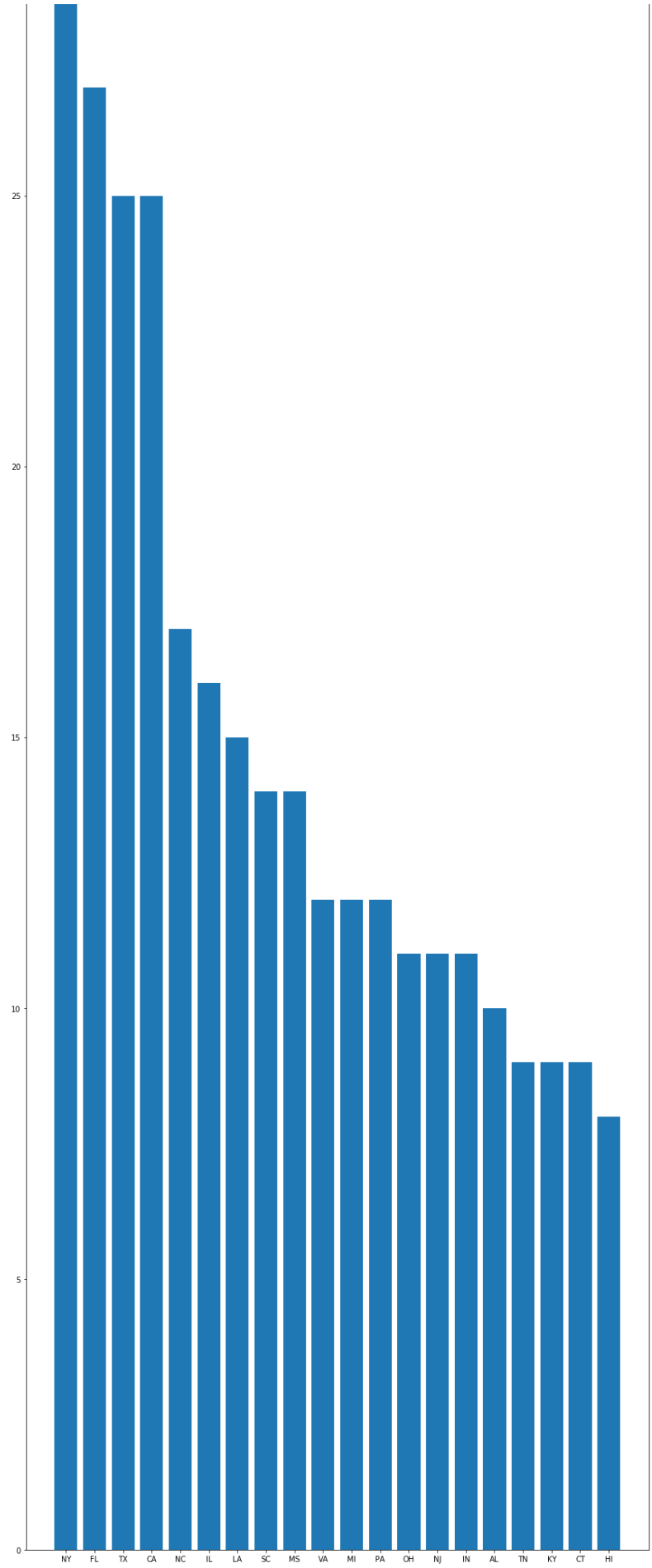
In [51]:

```
names = df['State'].iloc[:20]
values = df['City Participation Count'].iloc[:20]

plt.figure(figsize=(50, 40))

plt.subplot(131)
plt.bar(names, values)
plt.suptitle('Categorical Plotting')
plt.show()
```

Categorical Plotting



From the analysis, we can see New York has highest participation and our sweet home "Alabama" is in 16 th place regarding the number of participation. Cheers!