In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [ ]:
```
df = pd.read_csv('World University Rankings 2023.csv',encoding = 'iso-8859-1')
```

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2341 entries, 0 to 2340
Data columns (total 13 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   University Rank              2341 non-null   object 
 1   Name of University           2233 non-null   object 
 2   Location                     2047 non-null   object 
 3   No of student                2209 non-null   object 
 4   No of student per staff      2208 non-null   float64
 5   International Student        2209 non-null   object 
 6   Female:Male Ratio            2128 non-null   object 
 7   OverAll Score                1799 non-null   object 
 8   Teaching Score               1799 non-null   float64
 9   Research Score               1799 non-null   float64
 10  Citations Score              1799 non-null   float64
 11  Industry Income Score        1799 non-null   float64
 12  International Outlook Score  1799 non-null   float64
dtypes: float64(6), object(7)
memory usage: 237.9+ KB
```

In [ ]:
```
df.head()
```

Out[ ]:

	University Rank	Name of University	Location	No of student	No of student per staff	International Student	Female:Male Ratio	OverA Sco
0	1	University of Oxford	United Kingdom	20,965	10.6	42%	48 : 52	96
1	2	Harvard University	United States	21,887	9.6	25%	50 : 50	95
2	3	University of Cambridge	United Kingdom	20,185	11.3	39%	47 : 53	94
3	3	Stanford University	United States	16,164	7.1	24%	46 : 54	94
4	5	Massachusetts Institute of Technology	United States	11,415	8.2	33%	40 : 60	94

In [ ]:

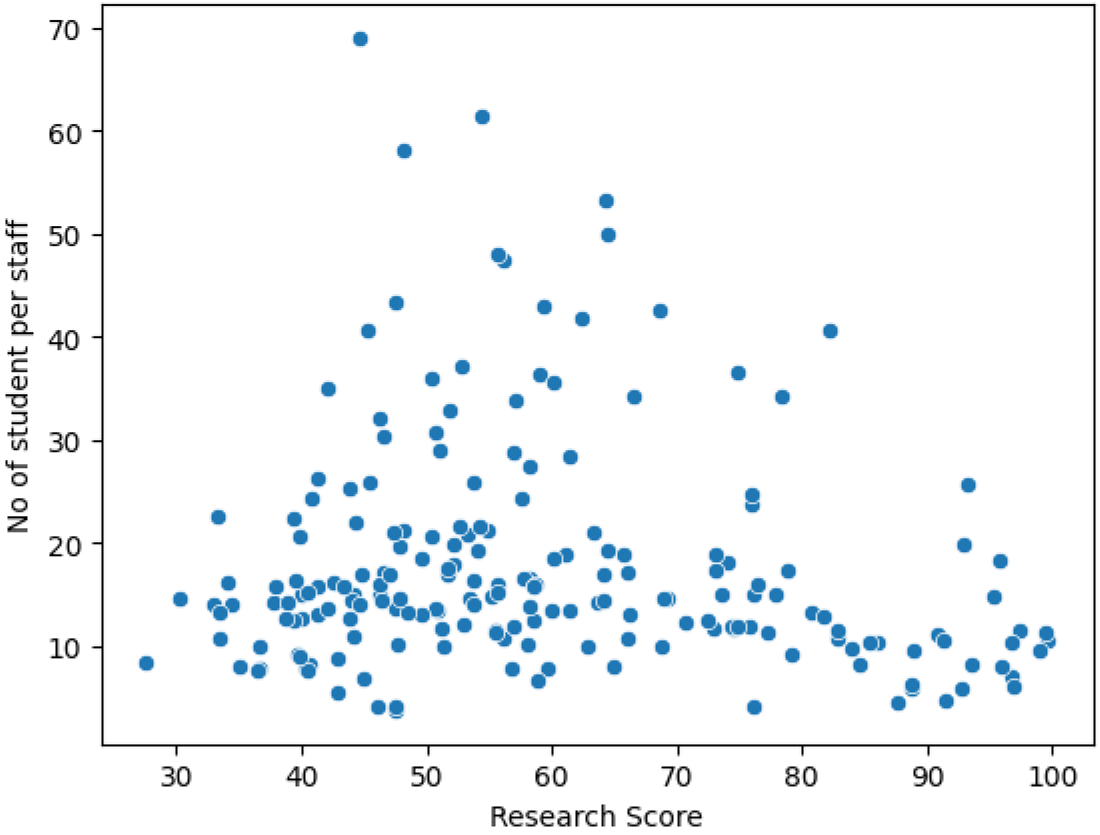
```
df2 = df[0:199]
```

In [ ]:

```
sns.scatterplot(data = df2 , x = 'Research Score', y = 'No of student per staff')
```

Out[ ]:

```
<Axes: xlabel='Research Score', ylabel='No of student per staff'>
```



In [ ]:

```
df3 = df2[["Research Score","No of student per staff"]].dropna()
```

In [ ]:

```
from sklearn.cluster import KMeans
```

In [ ]:

```
model = KMeans(n_clusters=4 , random_state=0)
model.fit(df3)
```

Out[ ]:

```
▾ KMeans
KMeans(n_clusters=4, random_state=0)
```

In [ ]:

```
model.cluster_centers_
```

Out[ ]:

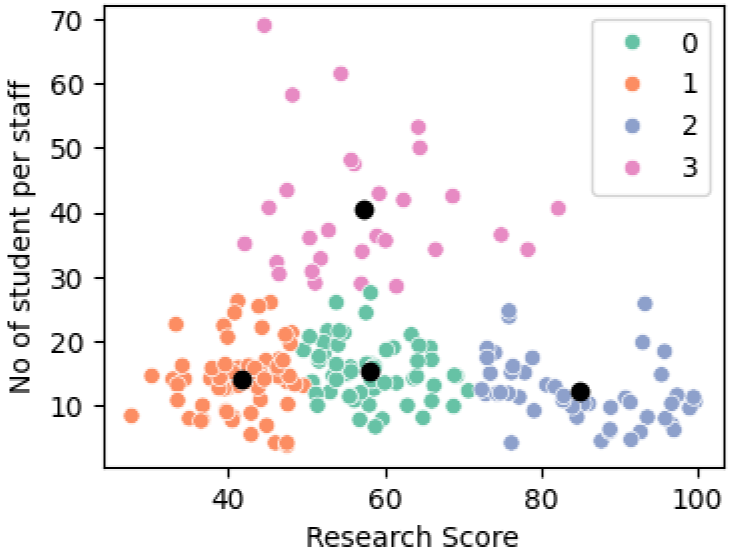
```
array([[58.11355932, 15.4559322 ],
       [41.54761905, 14.03650794],
       [84.87708333, 12.20416667],
       [57.21034483, 40.38275862]])
```

In [ ]:

```
plt.figure(figsize=[4,3])
sns.scatterplot(data = df3 , x = 'Research Score', y = 'No of student per staff'
                ,hue=model.labels_,palette='Set2')
```

```
plt.scatter(model.cluster_centers_[:,0], model.cluster_centers_[:,1]
            ,color = 'k',marker='o')
```

Out[ ]: `<matplotlib.collections.PathCollection at 0x1fe25f353d0>`



In [ ]:

```
model.predict([[12,93],[20,64]])
```

```
c:\Users\HP\Desktop\DataSci\.venv\Lib\site-packages\sklearn\base.py:493: UserWarnin
g: X does not have valid feature names, but KMeans was fitted with feature names
  warnings.warn(
```

Out[ ]: `array([3, 3])`