In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
```

In [2]:

```
df = pd.read_csv("Kmeans data.csv")
```

In [3]:

```
df.head()
```

Out[3]:

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

In [4]:

```
df.shape
```

Out[4]:

```
(200, 5)
```

In [5]:

```
df.columns
```

Out[5]:

```
Index(['CustomerID', 'Genre', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)'],
      dtype='object')
```

In [6]:

```
df1 = df.drop(columns=['CustomerID', 'Genre','Age'])
```

In [7]:

```
df1.head()
```

Out[7]:

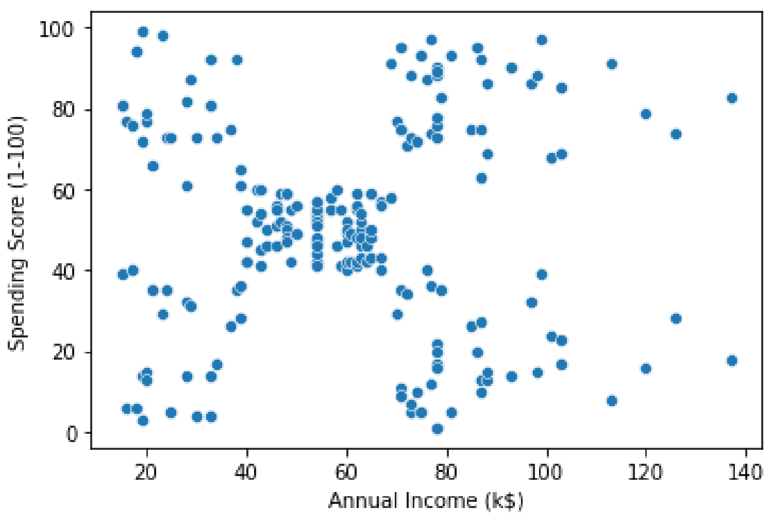
	Annual Income (k$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

In [8]:

```
sns.scatterplot(data = df1, x='Annual Income (k$)',y='Spending Score (1-100)')
```

Out[8]:

```
<AxesSubplot:xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'>
```



In [9]:

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

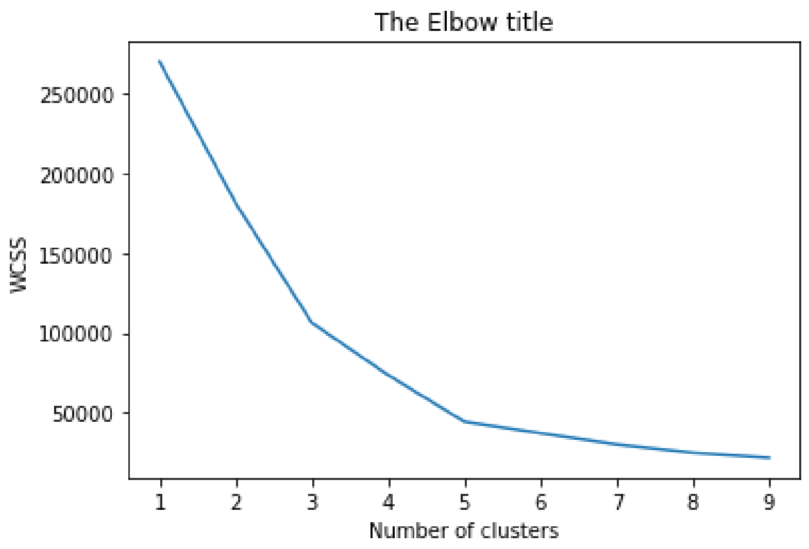
In [10]:

```
wcss = []
for i in range(1,10):
    kmeans = KMeans(i)
    kmeans.fit(df1)
    wcss.append(kmeans.inertia_)

number_clusters = range(1,10)
plt.plot(number_clusters,wcss)
plt.title('The Elbow title')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
```

Out[10]:

```
Text(0, 0.5, 'WCSS')
```



In [11]:

```
# selecting no. of clusters to be 5
Kmeans = KMeans(5)
Kmeans.fit(df1)
```

Out[11]:

```
KMeans(n_clusters=5)
```

In [12]:

```
Kmeans.labels_
```

Out[12]:

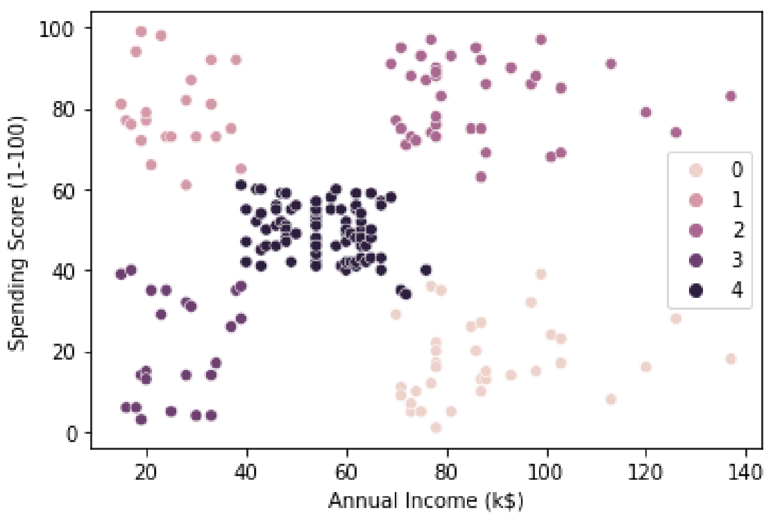
```
array([3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1,
       3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 4,
       3, 1, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 2, 0, 2, 4, 2, 0, 2, 0, 2,
       4, 2, 0, 2, 0, 2, 0, 2, 0, 2, 4, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,
       0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,
       0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,
       0, 2])
```

In [13]:

```
# Let's Validate how the clustering was done
sns.scatterplot(data = df1, x='Annual Income (k$)',y='Spending Score (1-100)', hue=Kmeans.labels_)
```

Out[13]:

```
<AxesSubplot:xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'>
```



Result

1. We have successfully clustered, and have a good number of clusters.
2. As it is an unsupervised algorithm we can't tell how well it might perform for more than two features, in such cases dimensinality reduction can help to visualize clusters.

In [ ]: