In [1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
df = pd.read_csv("C:/Users/ZJU/Downloads/Mall_Customers.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	CustomerID	Gender	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

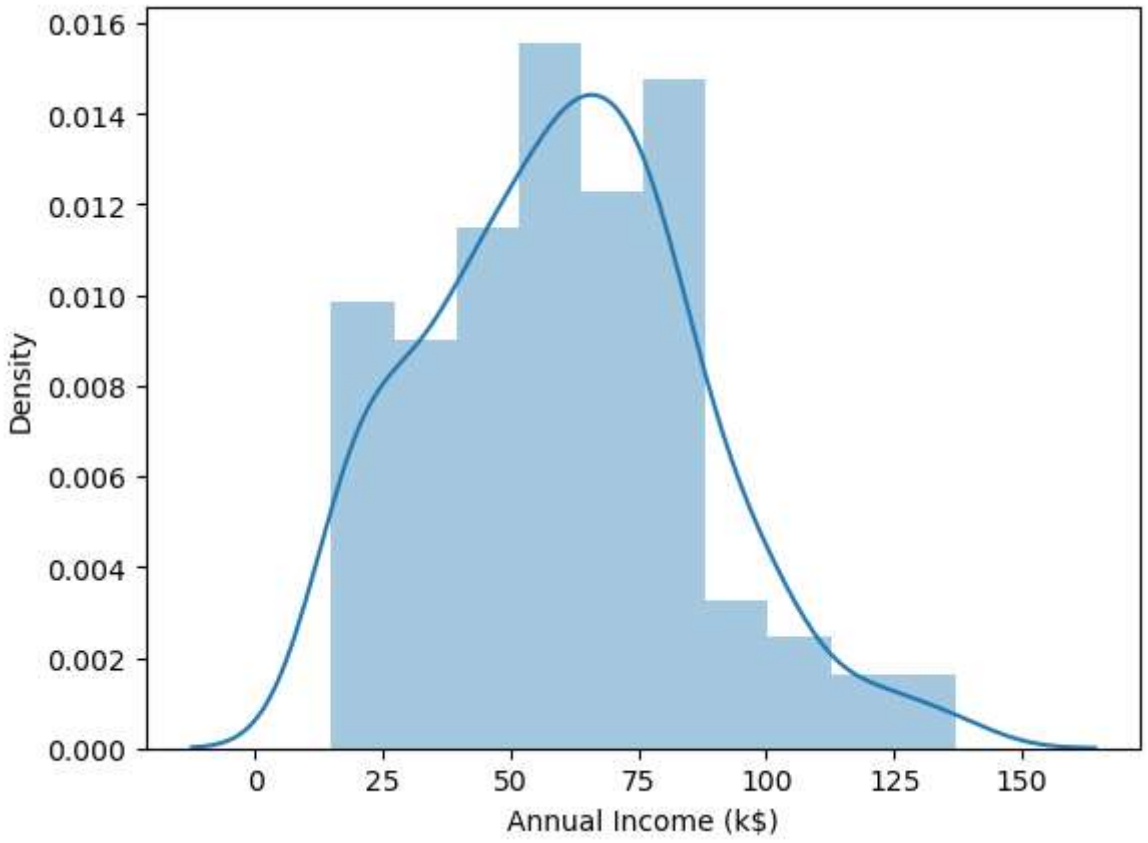
Univeriate Analysis

In [4]:
```
df.describe()
```

Out[4]:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

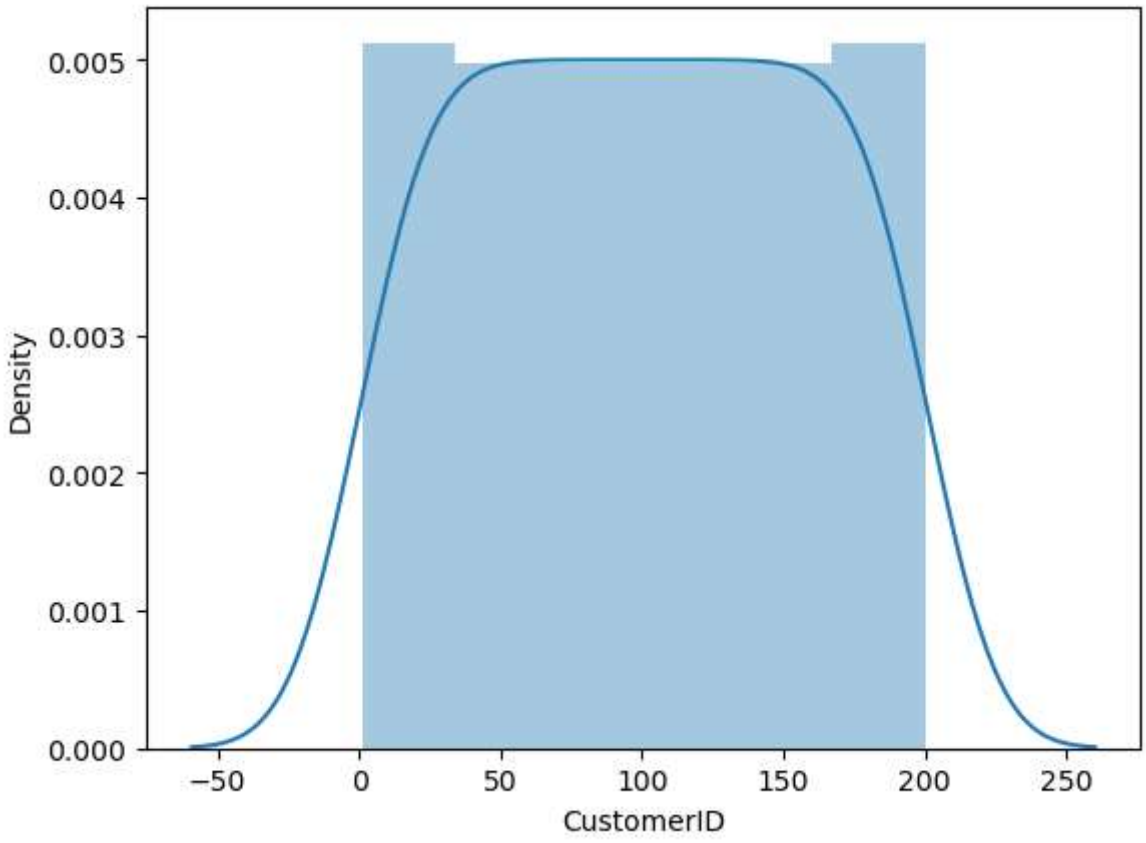
In [5]:
```
sns.distplot(df['Annual Income (k$)']);
```

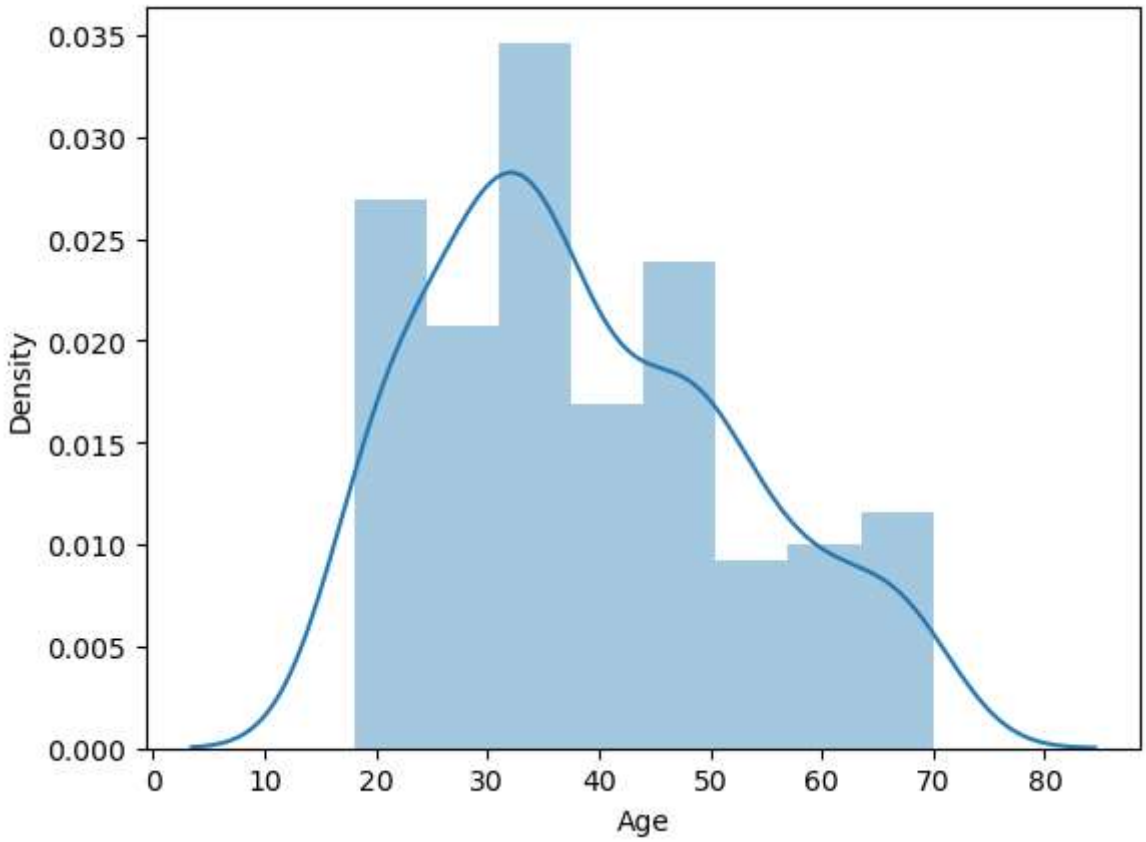


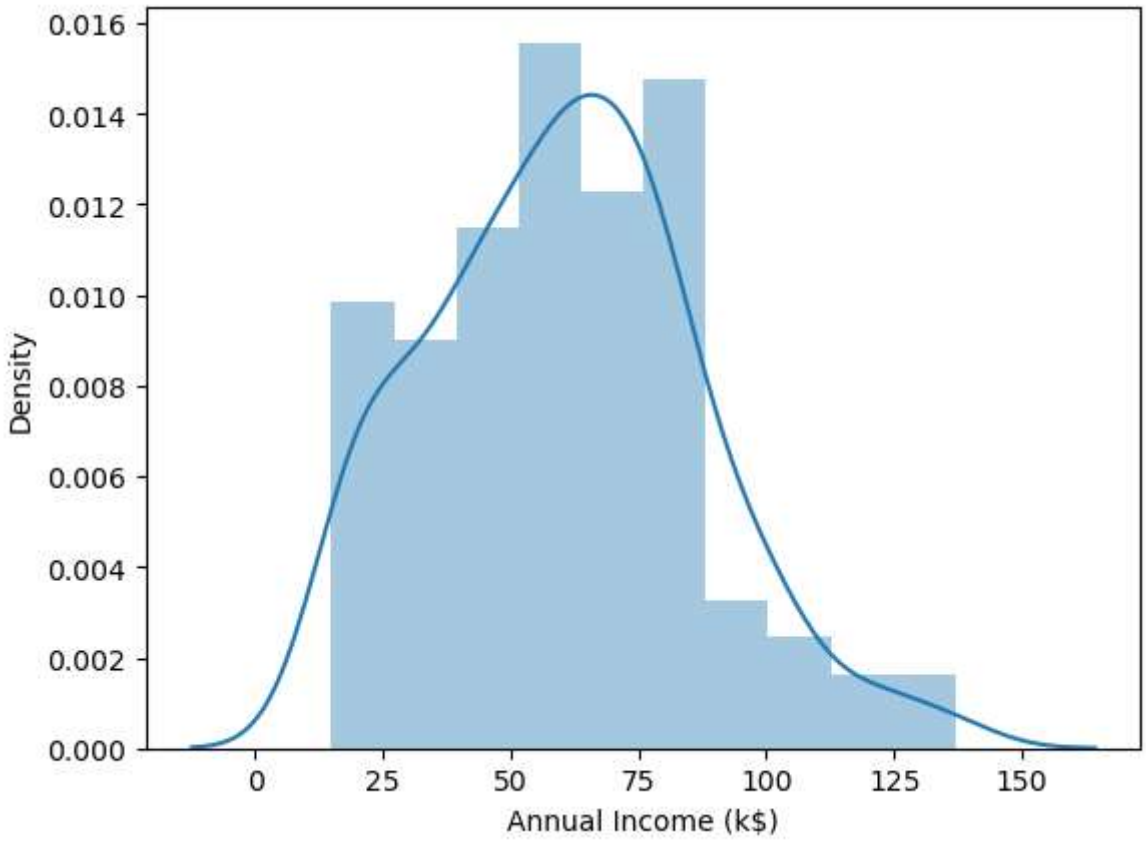
In [6]:
```
df.columns
```

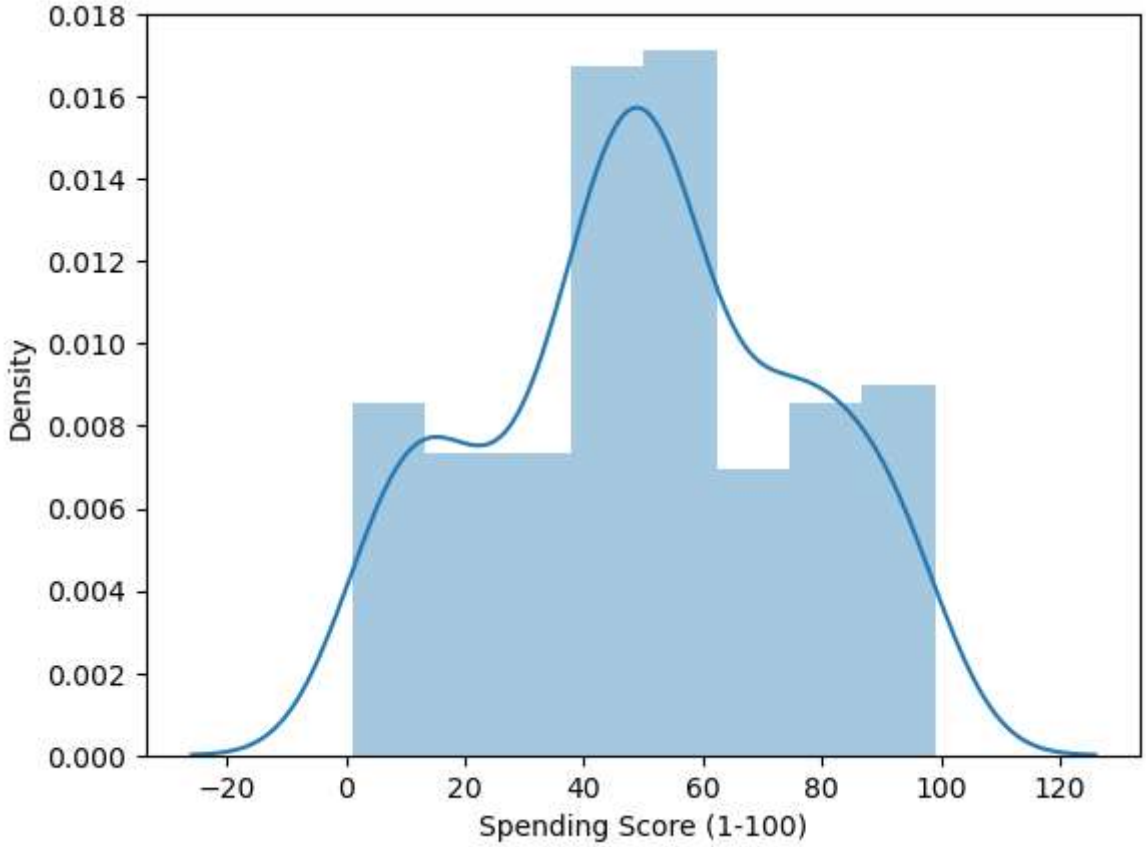
Out[6]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)'],
      dtype='object')
```

In [7]:
```
columns = ['CustomerID', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.distplot(df[i])
```

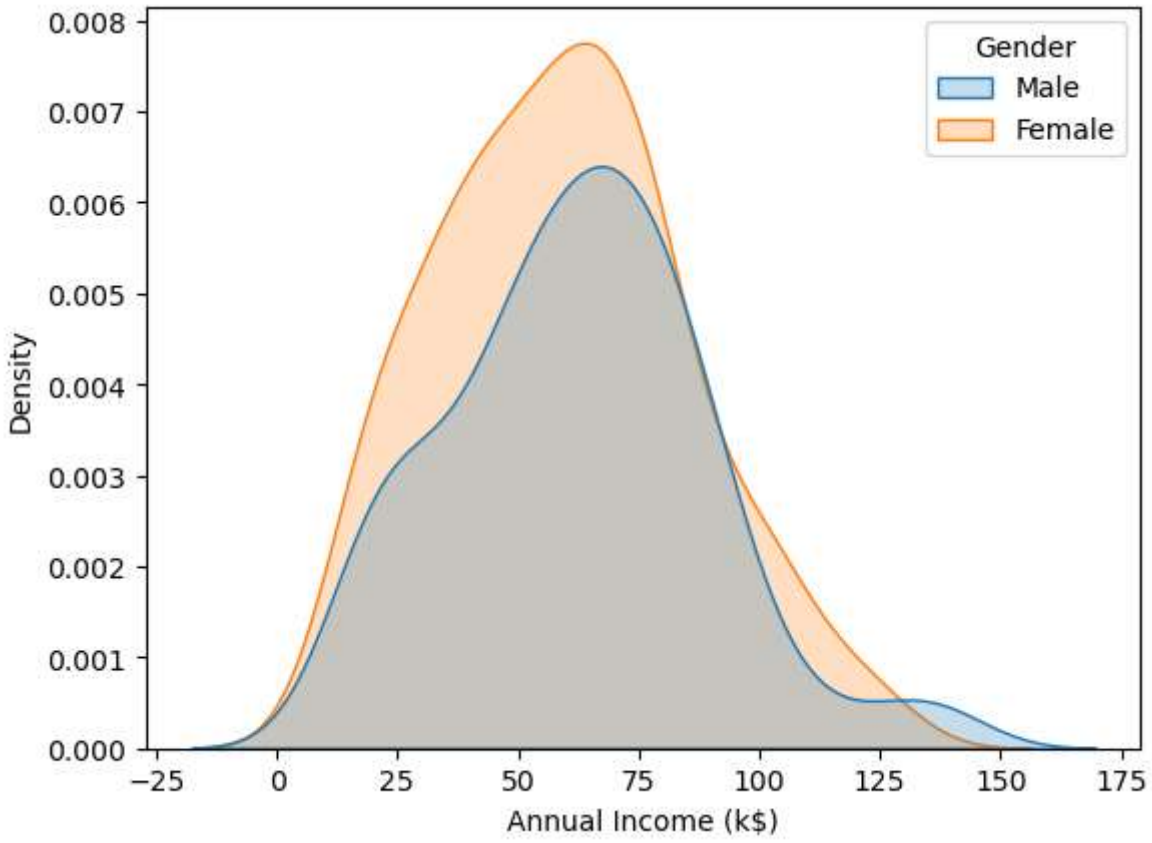

0.005
0.004
0.003
0.002
0.001
0.000
Density
−50
0
50
100
150
200
250
CustomerID


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
10
20
30
40
50
60
70
80
Age

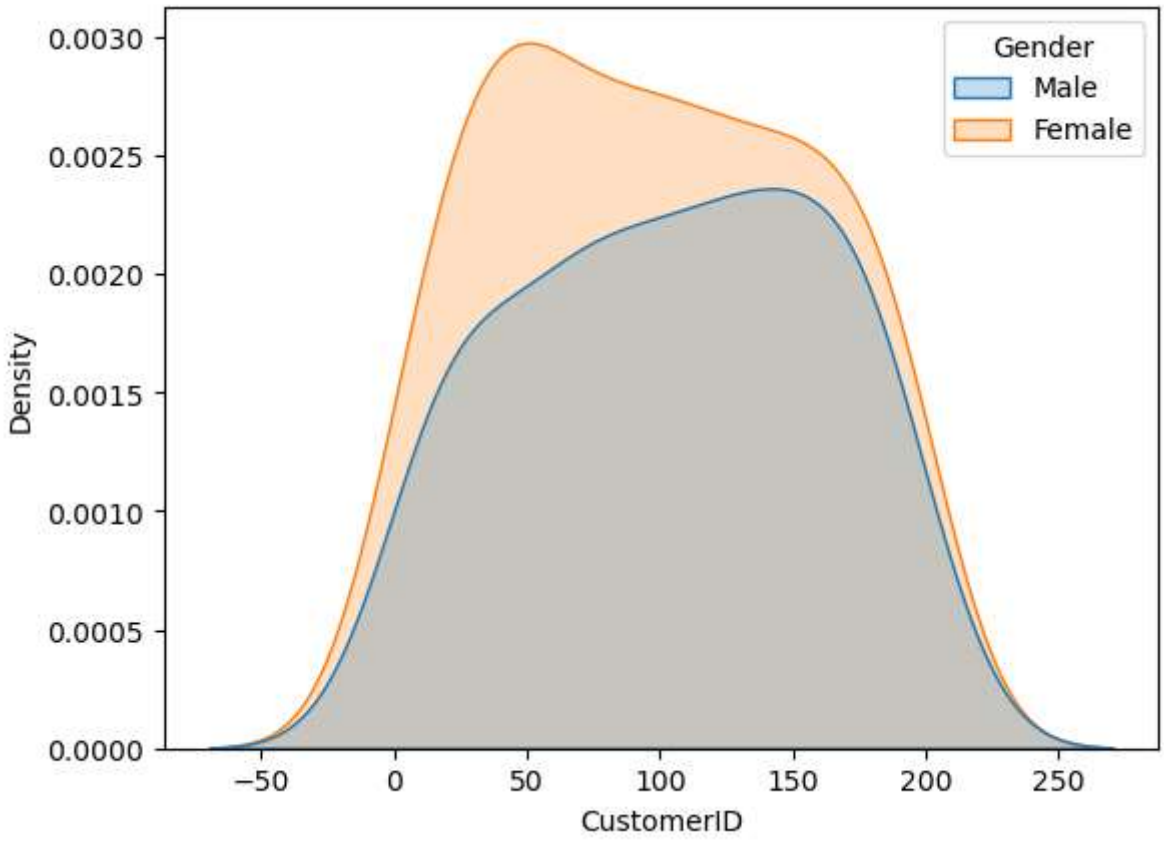


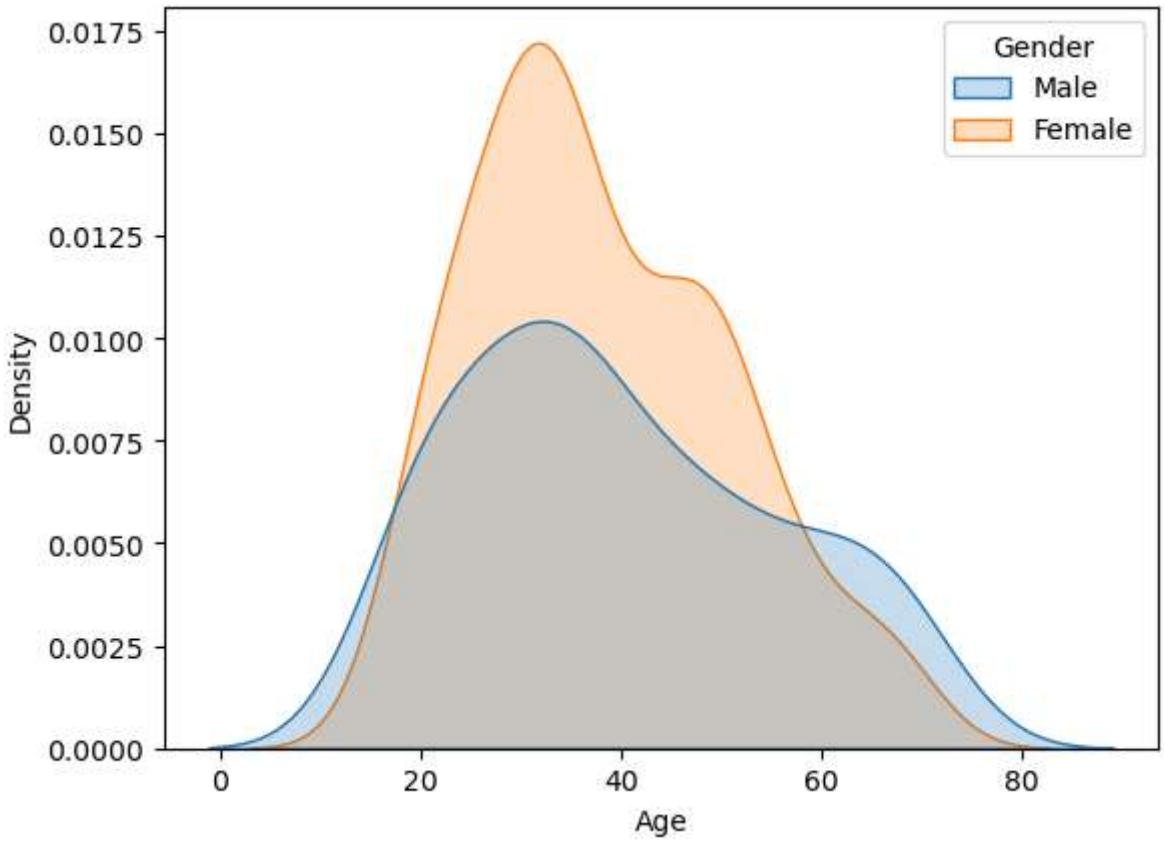


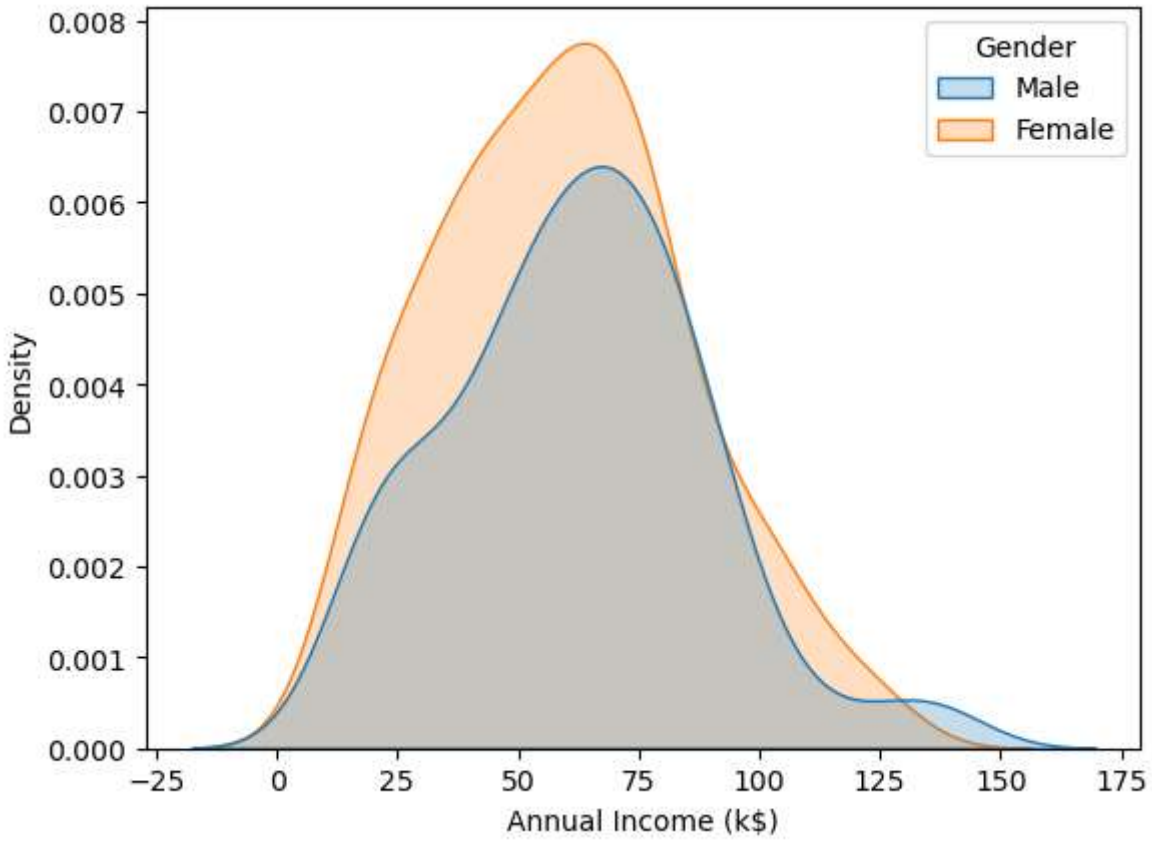
In [8]:
```
sns.kdeplot(x=df['Annual Income (k$)'],color='blue',shade=True,hue=df['Gender']);
```

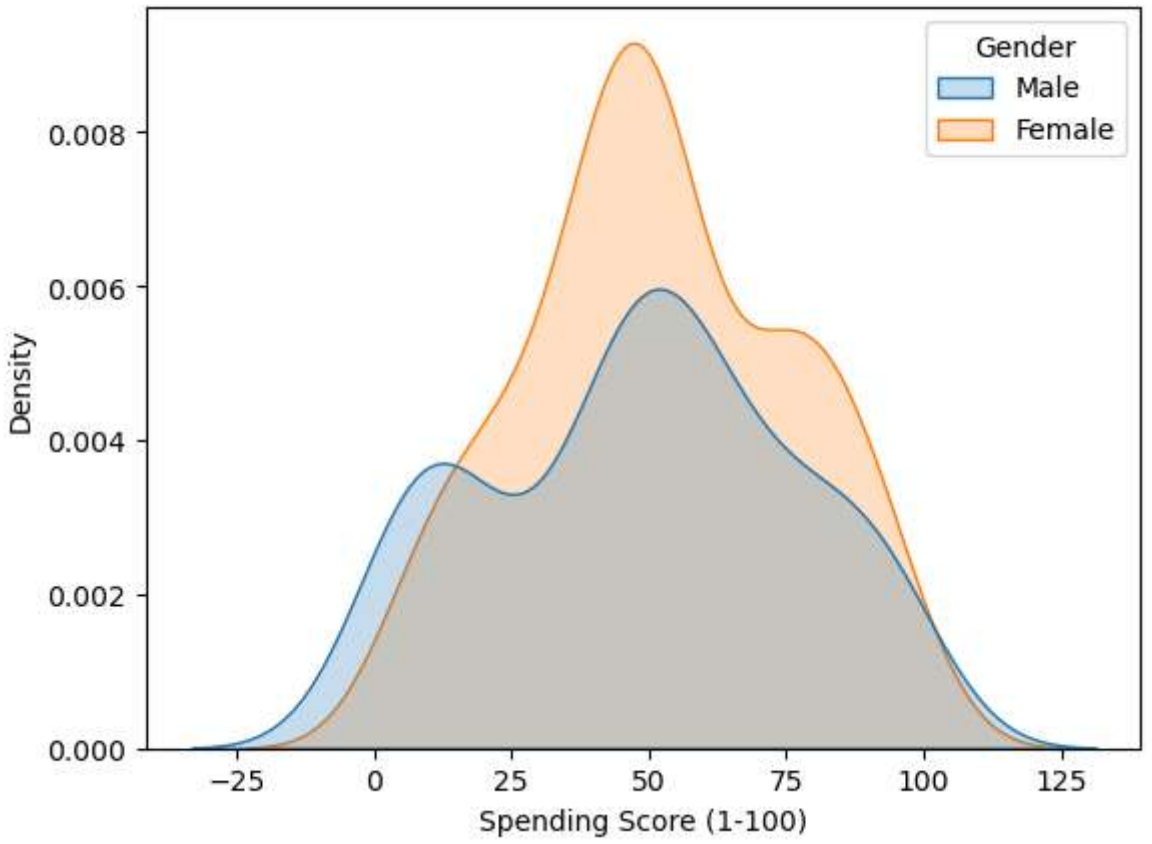


In [9]:
```
columns = ['CustomerID', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.kdeplot(x=df[i],shade=True,hue=df['Gender'],color="Green");
```

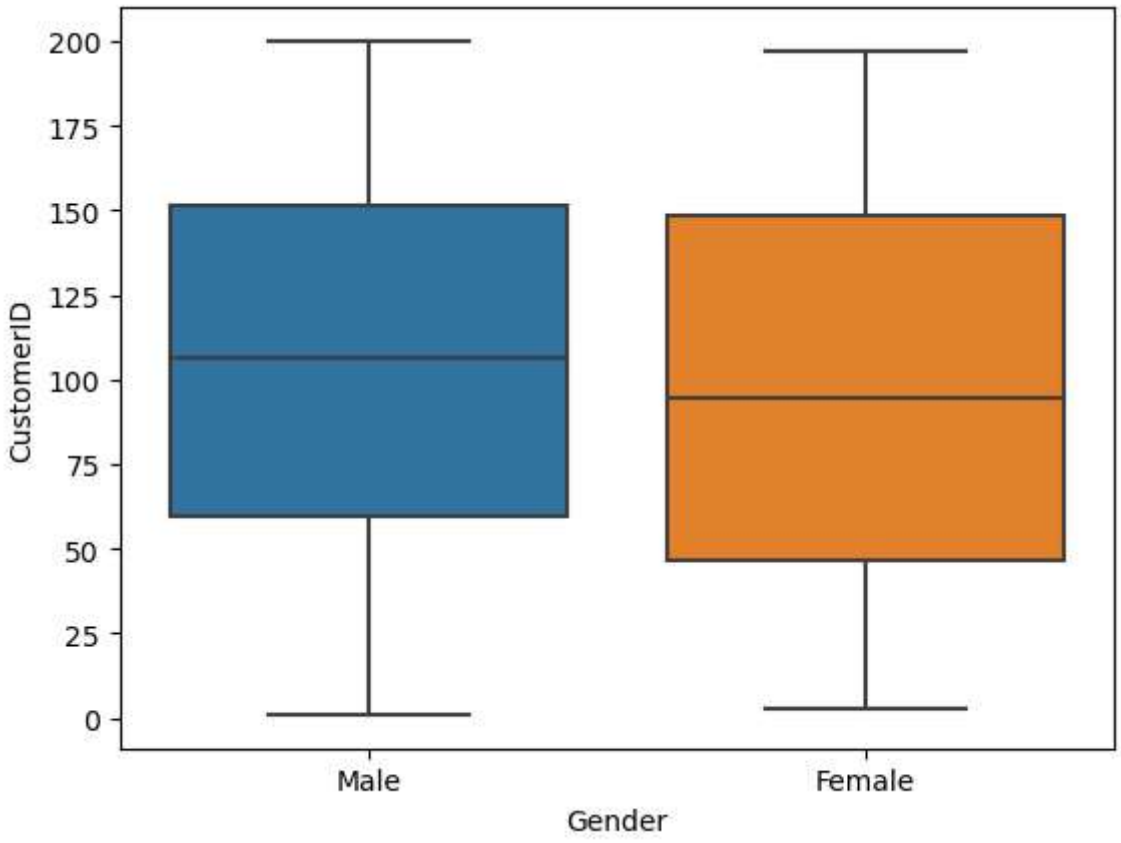

Gender
Male
Female
Density
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
−50
0
50
100
150
200
250
CustomerID

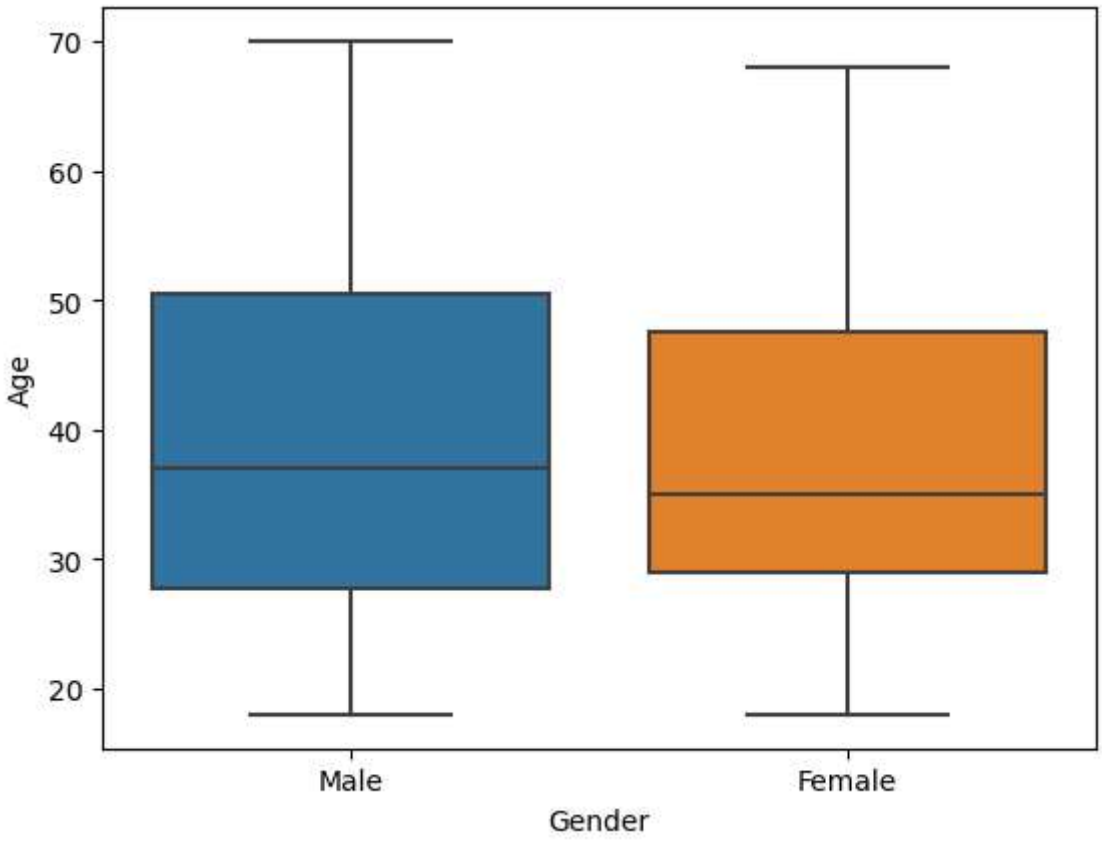

Gender
Male
Female
Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
20
40
60
80
Age

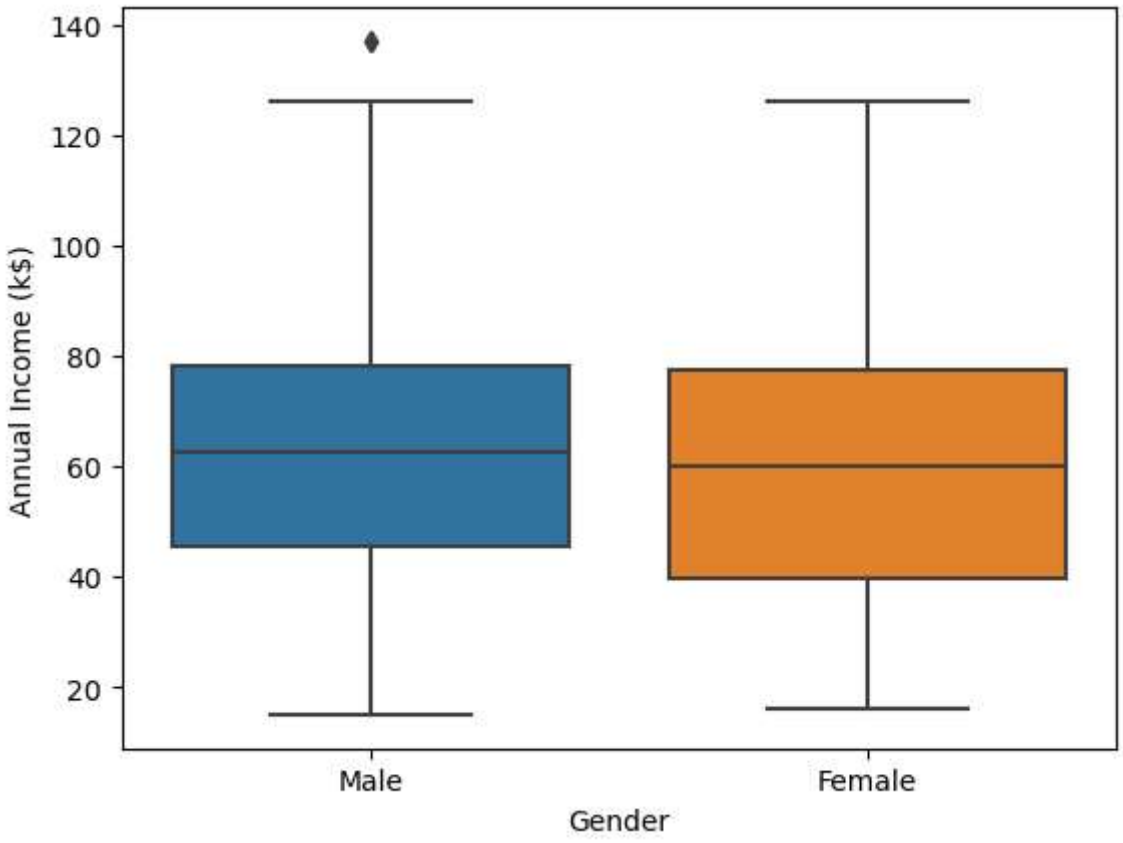


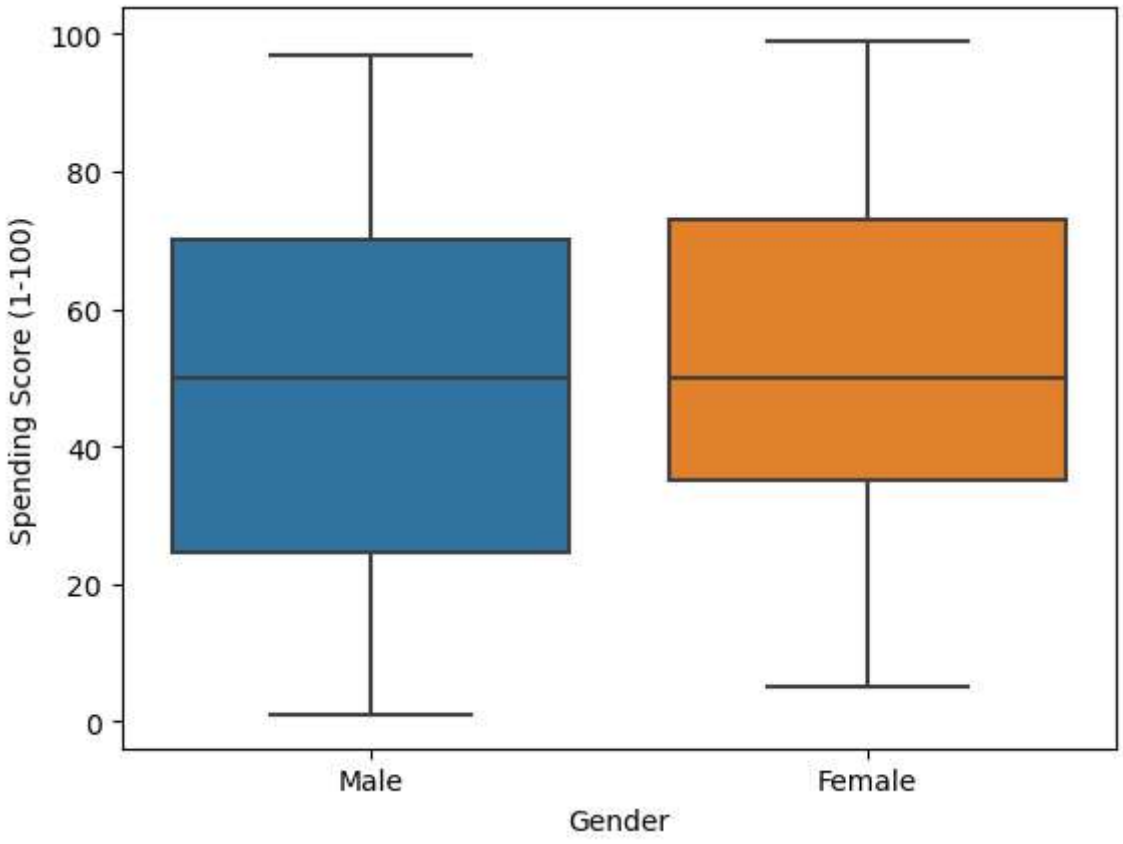


In [10]:
```
columns = ['CustomerID', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.boxplot(data=df, x='Gender',y=df[i])
```


200
175
150
125
100
75
50
25
0
CustomerID
Male
Female
Gender


70
60
50
40
30
20
Age
Male
Female
Gender





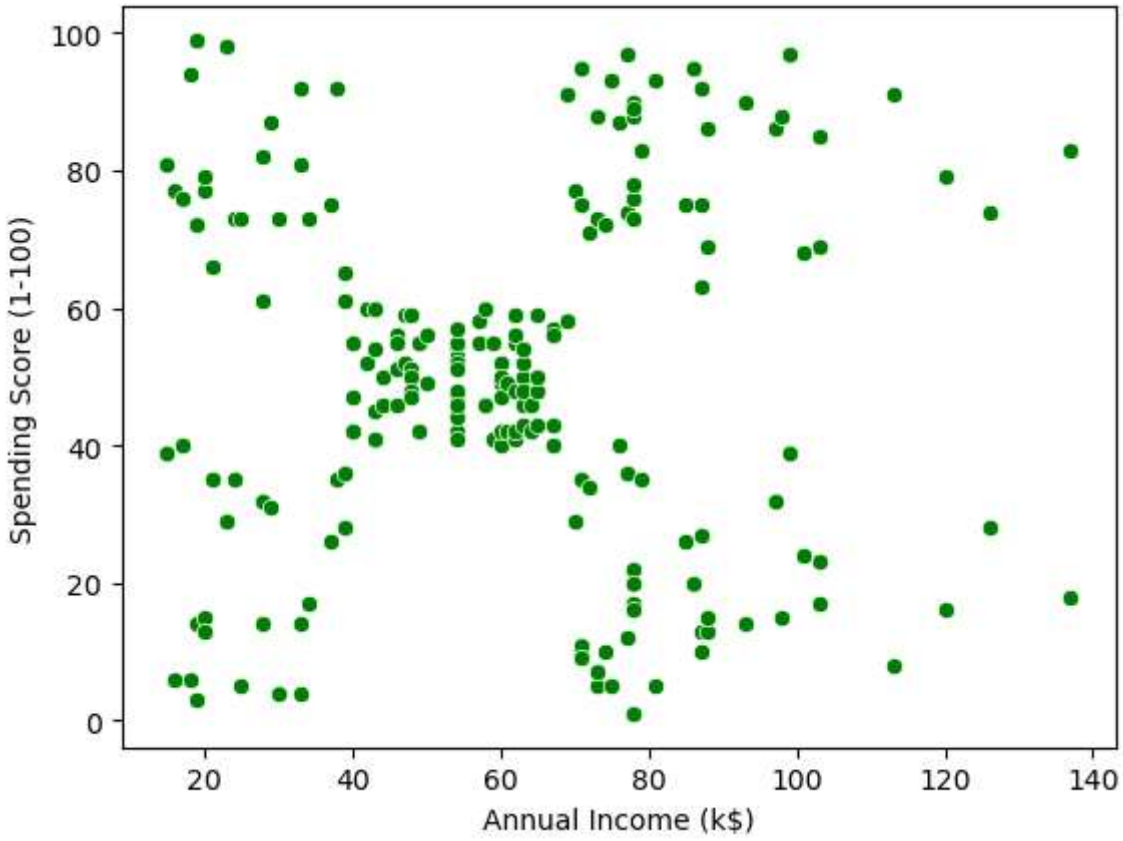
In [11]:
```
df['Gender'].value_counts()
```

Out[11]:
```
Female    112
Male       88
Name: Gender, dtype: int64
```

Bivariate Analysis

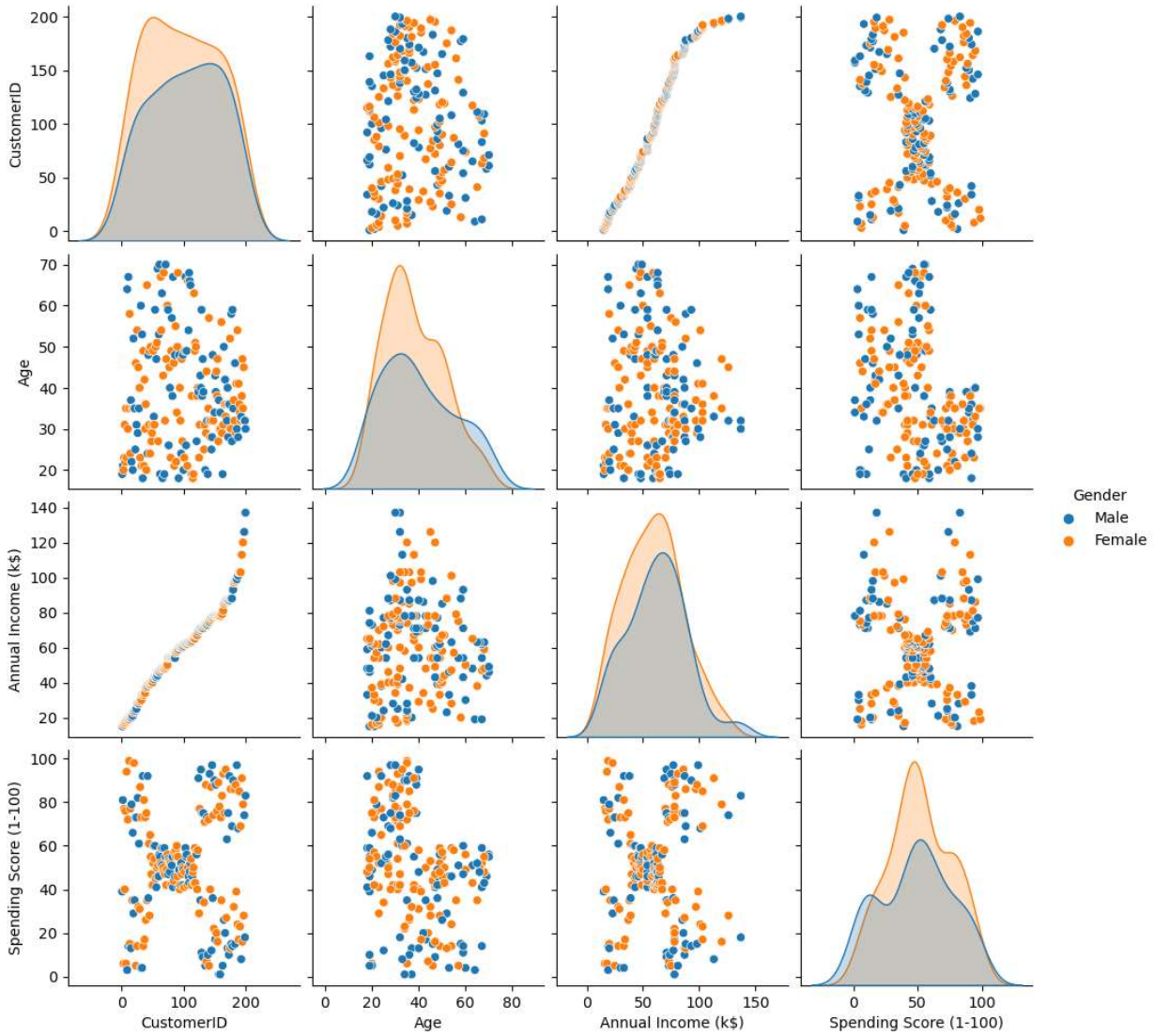
In [12]:
```
sns.scatterplot(data=df,x='Annual Income (k$)',y='Spending Score (1-100)',color='greer
```

Out[12]:
```
<Axes: xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'>
```



In [13]:
```
#df=df.drop('CustomerID',axis=1)
sns.pairplot(df,hue='Gender')
```

Out[13]:
```
<seaborn.axisgrid.PairGrid at 0x23b4c634a10>
```



In [14]:
```
df.groupby(['Gender'])['Age', 'Annual Income (k$)',
       'Spending Score (1-100)'].mean()
```

Out[14]:

Gender	Age	Annual Income (k$)	Spending Score (1-100)
Female	38.098214	59.250000	51.526786
Male	39.806818	62.227273	48.511364

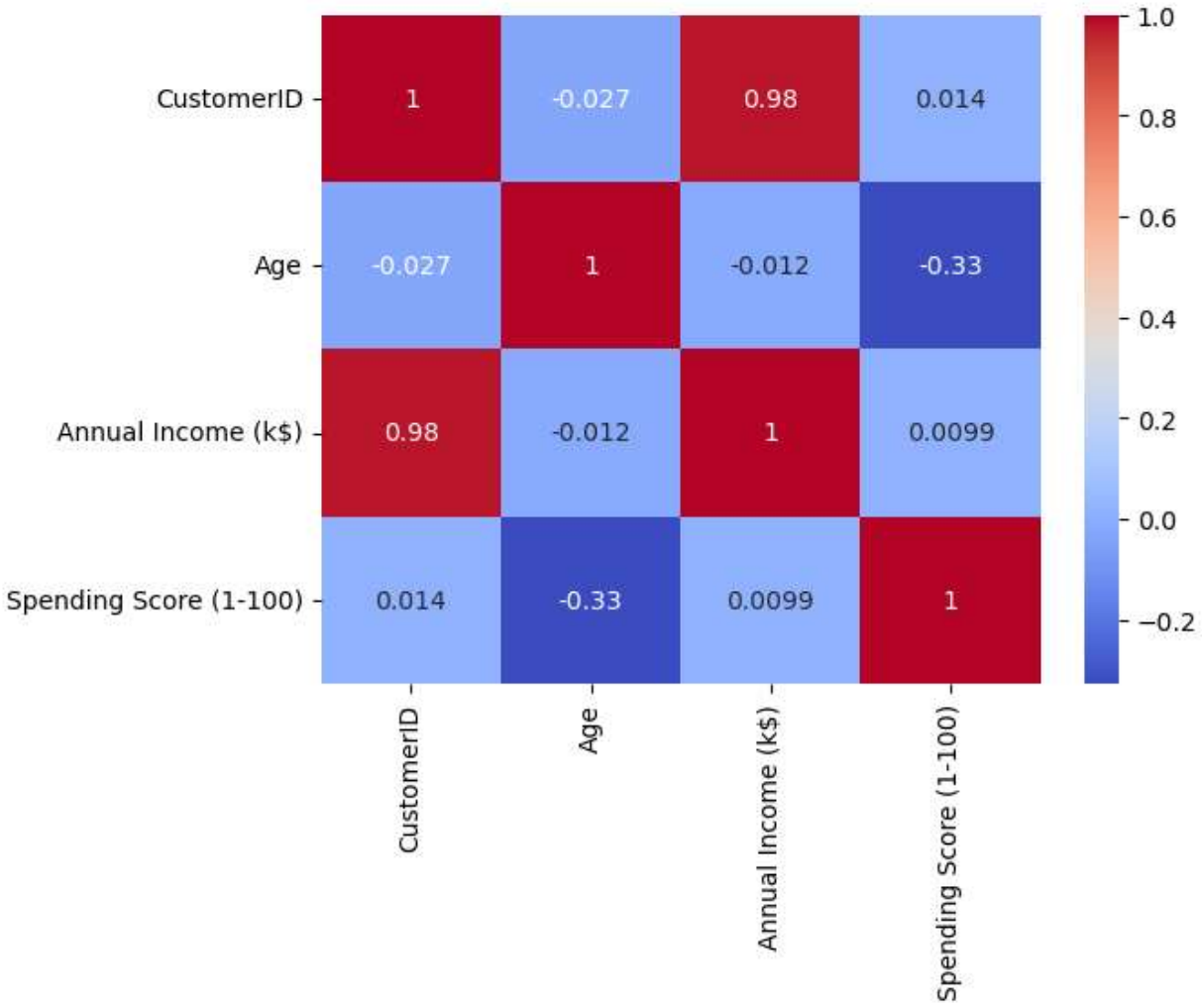
In [15]:
```
df.corr()
```

Out[15]:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
CustomerID	1.000000	-0.026763	0.977548	0.013835
Age	-0.026763	1.000000	-0.012398	-0.327227
Annual Income (k$)	0.977548	-0.012398	1.000000	0.009903
Spending Score (1-100)	0.013835	-0.327227	0.009903	1.000000

In [16]:
```
sns.heatmap(df.corr(),annot=True,cmap='coolwarm')
```

Out[16]: <Axes: >



Clustering-Univariate,Bivariate, Multivariate

In [17]:
```
clustering_1=KMeans(n_clusters=6)
```

In [18]:
```
clustering_1.fit(df[['Annual Income (k$)']])
```

Out[18]:
```
KMeans
KMeans(n_clusters=6)
```

In [19]:
```
clustering_1
```

Out[19]:
```
KMeans
KMeans(n_clusters=6)
```

In [20]:
```
clustering_1.labels_
```

Out[20]:
```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 5, 5, 5, 5, 5, 5, 5, 5,
       5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,
       5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 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, 2, 2, 2, 2,
       2, 2])
```

In [21]:
```
df['Income Cluster'] = clustering_1.labels_
```

In [22]:
```
df.head()
```

Out[22]:

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

In [23]:
```
df['Income Cluster'].value_counts()
```

Out[23]:
```
0    48
3    42
5    42
1    32
4    28
2     8
Name: Income Cluster, dtype: int64
```

In [24]:
```
clustering_1.inertia_
```

Out[24]:
```
5050.904761904762
```

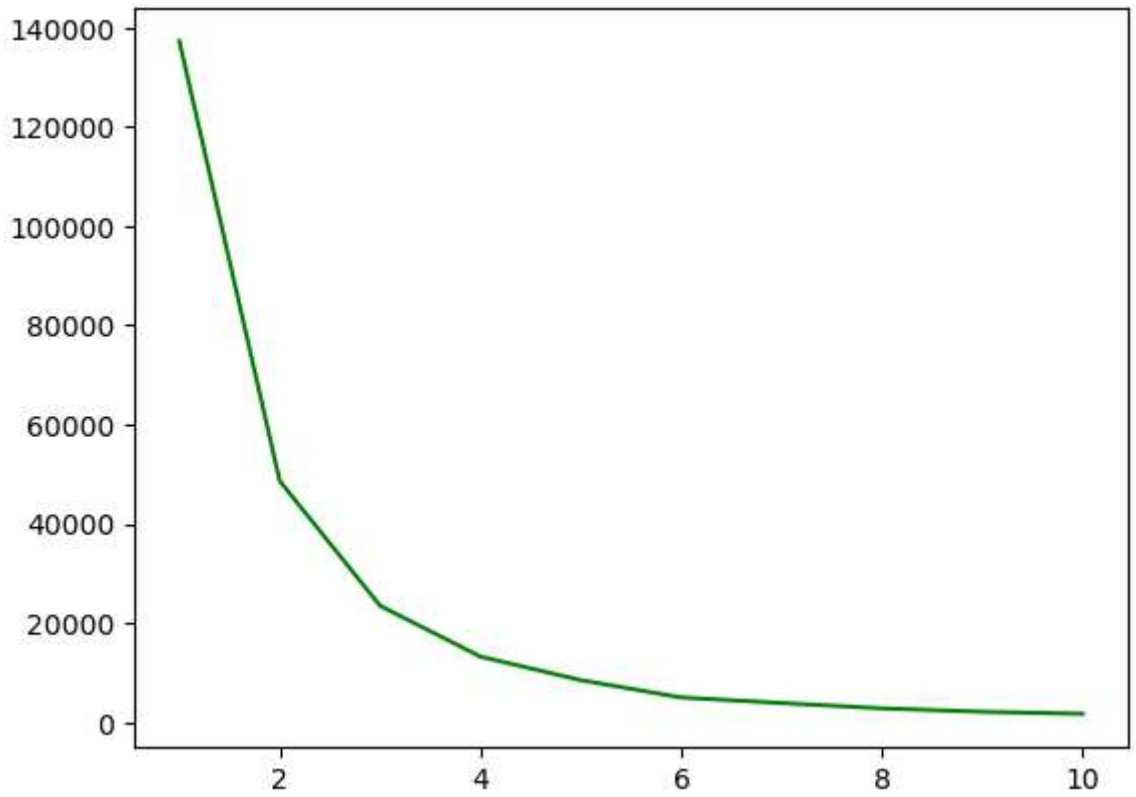
In [25]:
```
inertia_scores=[]
for i in range (1,11):
    kmeans=KMeans(n_clusters=i)
    kmeans.fit(df[['Annual Income (k$)']])
    inertia_scores.append(kmeans.inertia_)
```

In [26]:
```
inertia_scores
```

Out[26]:
```
[137277.27999999997,
 48660.88888888889,
 23517.330930930933,
 13278.112713472485,
 8534.415154553048,
 5050.904761904762,
 3931.988095238096,
 2857.4416971916976,
 2168.4787157287155,
 1758.1453823953823]
```

In [27]:
```
plt.plot(range(1,11),inertia_scores,color='green')
```

Out[27]:
```
[<matplotlib.lines.Line2D at 0x23b4cb23ad0>]
```



In [28]:
```
df.columns
```

Out[28]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)', 'Income Cluster'],
      dtype='object')
```

In [29]:
```
df.groupby('Income Cluster')['Age','Annual Income (k$)','Spending Score (1-100)'].mean
```

Out[29]:

Income Cluster	Age	Annual Income (k$)	Spending Score (1-100)
0	41.604167	60.083333	49.041667
1	34.906250	22.000000	49.656250
2	36.500000	124.000000	49.625000
3	43.000000	42.238095	50.666667
4	38.214286	93.000000	50.928571
5	35.428571	75.095238	51.095238

Bivariate Clustering

In [30]:
```
clustering_2=KMeans(n_clusters=5)
clustering_2.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
```

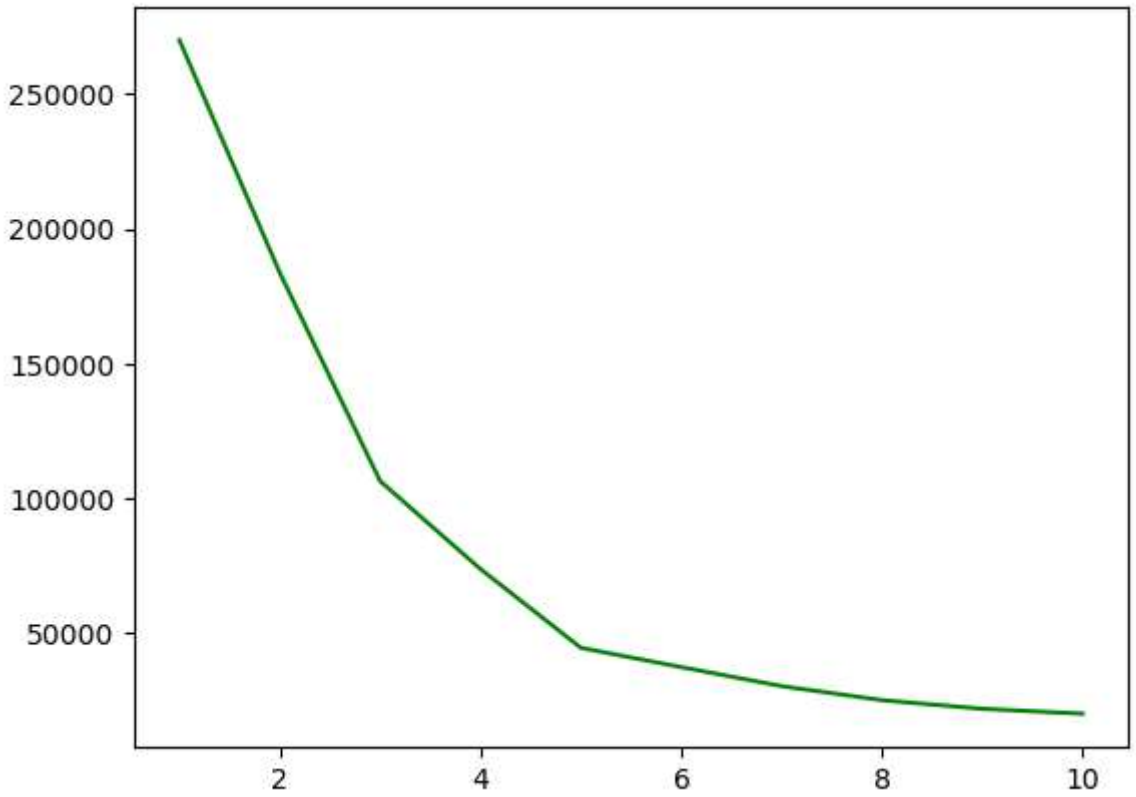
```
df['Spending and Income Clusters'] =clustering_2.labels_
df.head()
```

Out[30]:

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

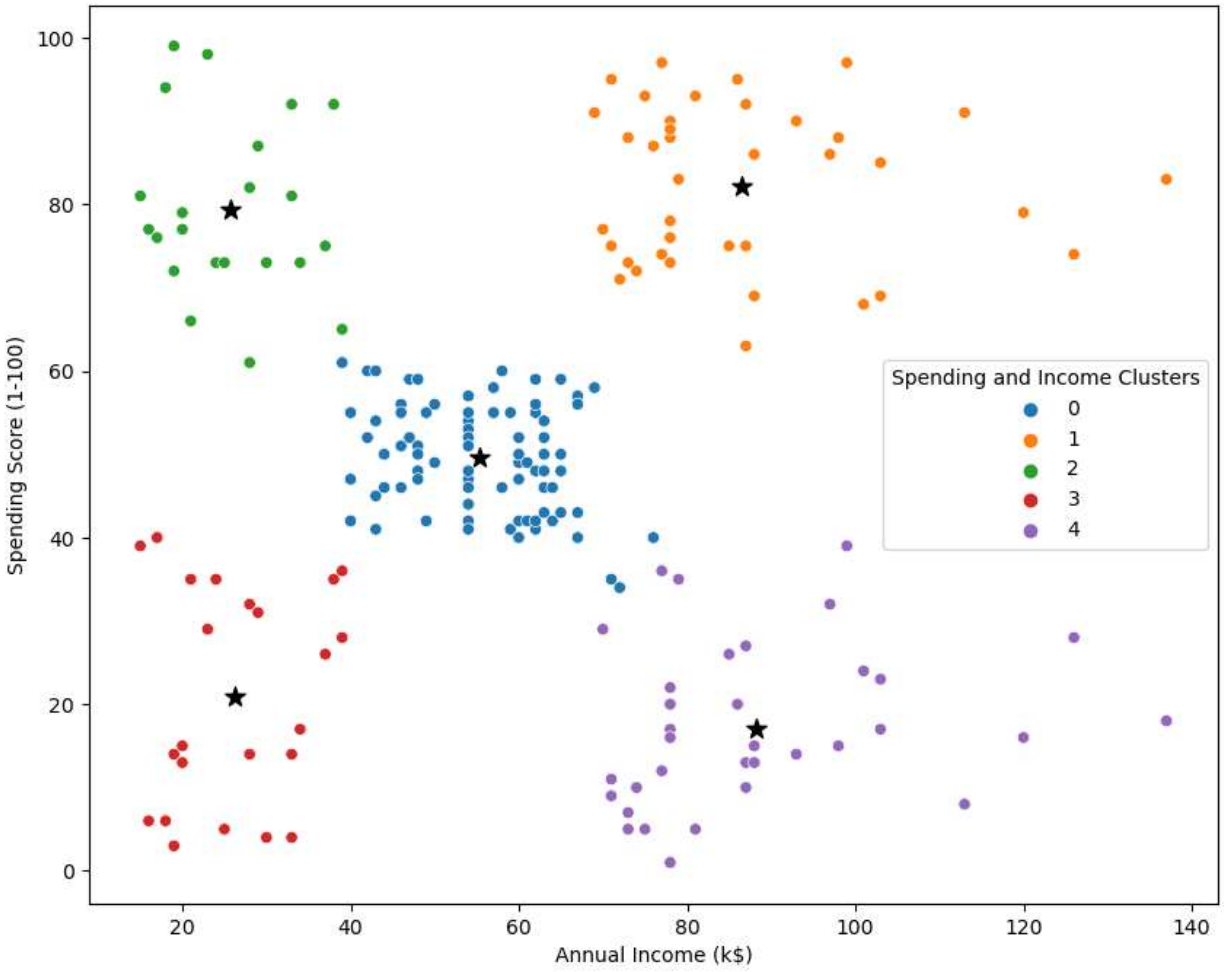
In [31]:
```
intertia_scores2=[]
for i in range(1,11):
    kmeans2=KMeans(n_clusters=i)
    kmeans2.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
    intertia_scores2.append(kmeans2.inertia_)
plt.plot(range(1,11),intertia_scores2,color='green')
```

Out[31]:
```
[<matplotlib.lines.Line2D at 0x23b4c762450>]
```



In [33]:
```
centers =pd.DataFrame(clustering_2.cluster_centers_)
centers.columns = ['x','y']
```

In [37]:
```
plt.figure(figsize=(10,8))
plt.scatter(x=centers['x'],y=centers['y'],s=100,c='black',marker='*')
sns.scatterplot(data=df, x ='Annual Income (k$)',y='Spending Score (1-100)',hue="Spend
plt.savefig('clustering_bivaraiate.png')
```



In [38]:
```
pd.crosstab(df['Spending and Income Clusters'],df['Gender'],normalize='index')
```

Out[38]:

Spending and Income Clusters	Gender	Female	Male
	0	0.592593	0.407407
	1	0.538462	0.461538
	2	0.590909	0.409091
	3	0.608696	0.391304
	4	0.457143	0.542857

In [39]:
```
df.groupby('Spending and Income Clusters')['Age','Annual Income (k$)','Spending Score
```

Out[39]:

Spending and Income Clusters	Age	Annual Income (k$)	Spending Score (1-100)
0	42.716049	55.296296	49.518519
1	32.692308	86.538462	82.128205
2	25.272727	25.727273	79.363636
3	45.217391	26.304348	20.913043
4	41.114286	88.200000	17.114286

Multivariate Clsutering

In [47]:
```
from sklearn.preprocessing import StandardScaler
```

In [56]:
```
scale= StandardScaler()
```

In [57]:
```
df.head()
```

Out[57]:

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

In [58]:
```
dff = pd.get_dummies(df,drop_first=True)
dff.head()
```

Out[58]:

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

In [59]:
```
dff.columns
```

Out[59]:
```
Index(['CustomerID', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)',
       'Income Cluster', 'Spending and Income Clusters', 'Gender_Male'],
      dtype='object')
```

In [60]:
```
dff=dff[['Age', 'Annual Income (k$)', 'Spending Score (1-100)', 'Spending and Income C
dff.head()
```

Out[60]:

	Age	Annual Income (k$)	Spending Score (1-100)	Spending and Income Clusters	Gender_Male
0	19	15	39	3	1
1	21	15	81	2	1
2	20	16	6	3	0
3	23	16	77	2	0
4	31	17	40	3	0

In [61]:
```
dff = scale.fit_transform(dff)
```

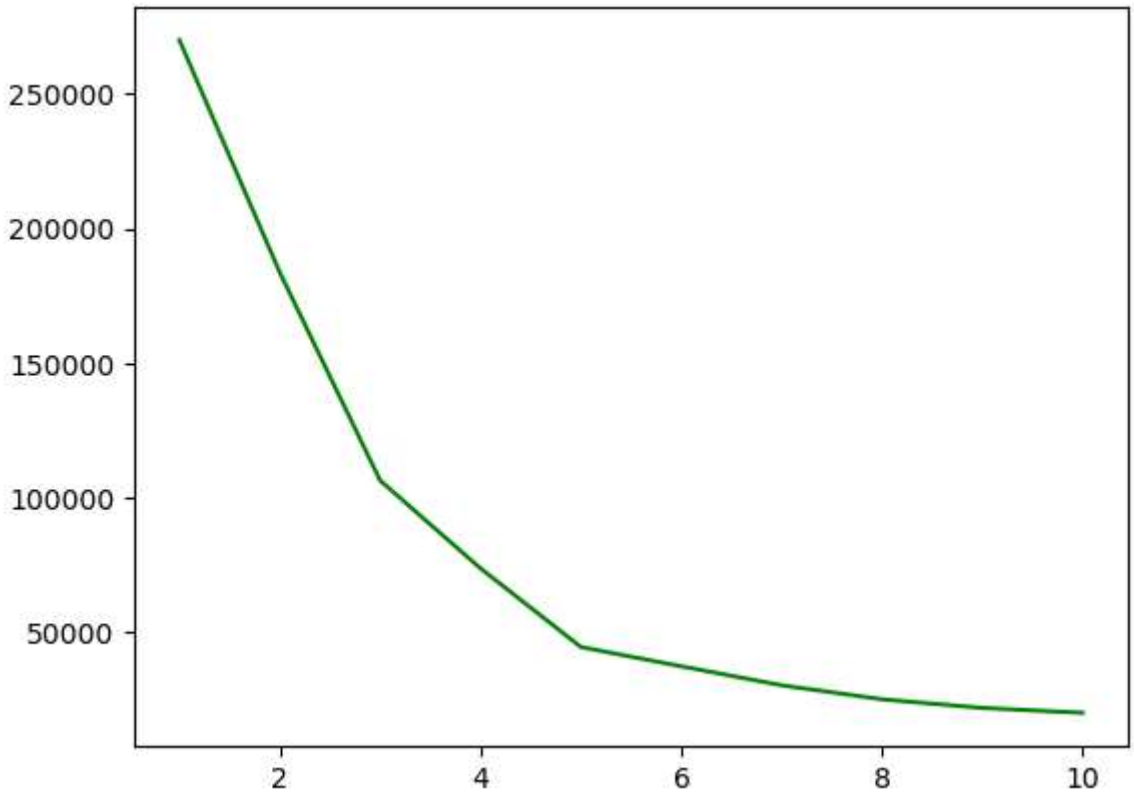
In [62]:
```
dff = pd.DataFrame(scale.fit_transform(dff))
dff.head()
```

Out[62]:

	0	1	2	3	4
0	-1.424569	-1.738999	-0.434801	1.007074	1.128152
1	-1.281035	-1.738999	1.195704	0.353130	1.128152
2	-1.352802	-1.700830	-1.715913	1.007074	-0.886405
3	-1.137502	-1.700830	1.040418	0.353130	-0.886405
4	-0.563369	-1.662660	-0.395980	1.007074	-0.886405

In [63]:
```
intertia_scores3=[]
for i in range(1,11):
    kmeans3=KMeans(n_clusters=i)
    kmeans3.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
    intertia_scores3.append(kmeans3.inertia_)
plt.plot(range(1,11),intertia_scores2,color='green')
```

Out[63]:
```
[<matplotlib.lines.Line2D at 0x23b4ef1b0d0>]
```



In [66]:
```
df.head()
```

Out[66]:

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

Conclusion

after clustering our analysis find out Target Cluster

1) Traget group would be cluster 1 which has a high Spending Score and high income: which means there is highly likelyhood of continue spending more narket team should foucs on this cluster

2) 53 % of cluster 1 shoppers are females.We should look for ways to attract these customers using a marketing campaign targeting popular items in this cluster.

3) Cluster 2 presents an interesting opportunity to market to the customers for sales event on popular items because of three reasons a) high spending core with low annual income with

average age of 25 years old.

In [ ]: