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 [6]:
```
df = pd.read_csv("Mall_Customers.csv")
```

In [7]:
```
df.head()
```

Out[7]:

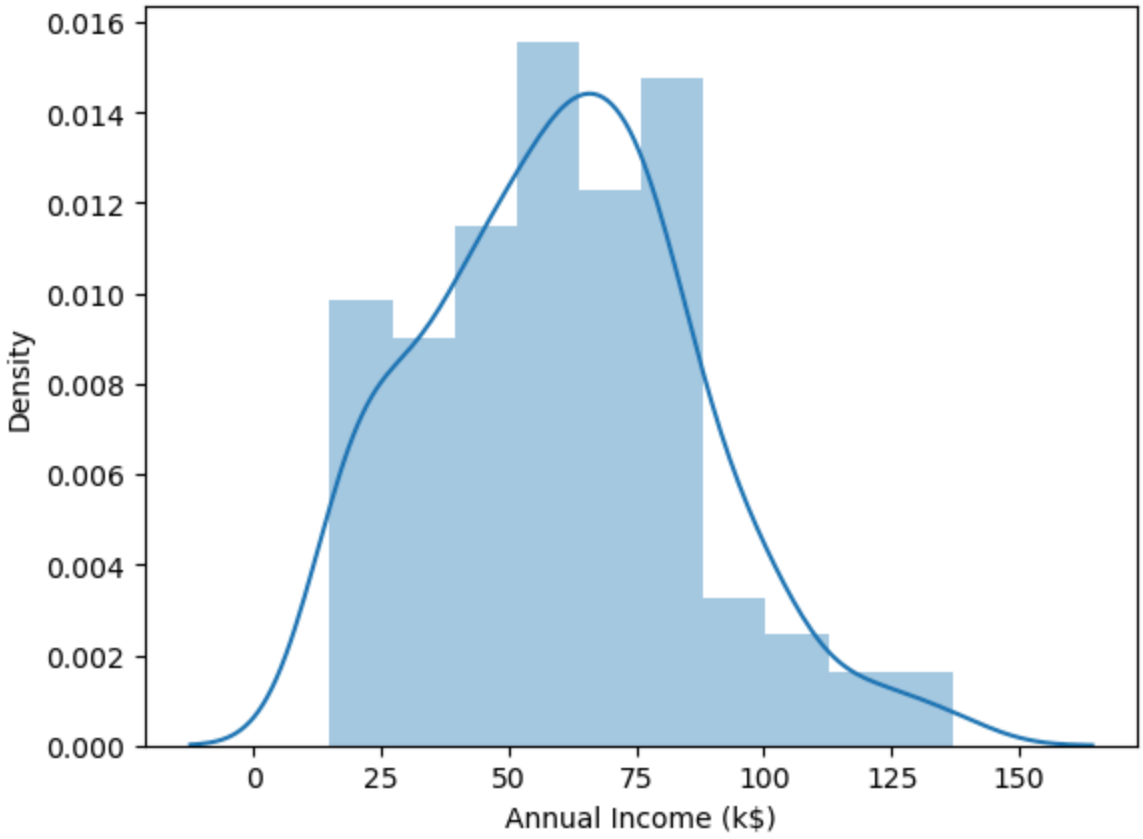
	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

In [8]:
```
df.describe()
```

Out[8]:

	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

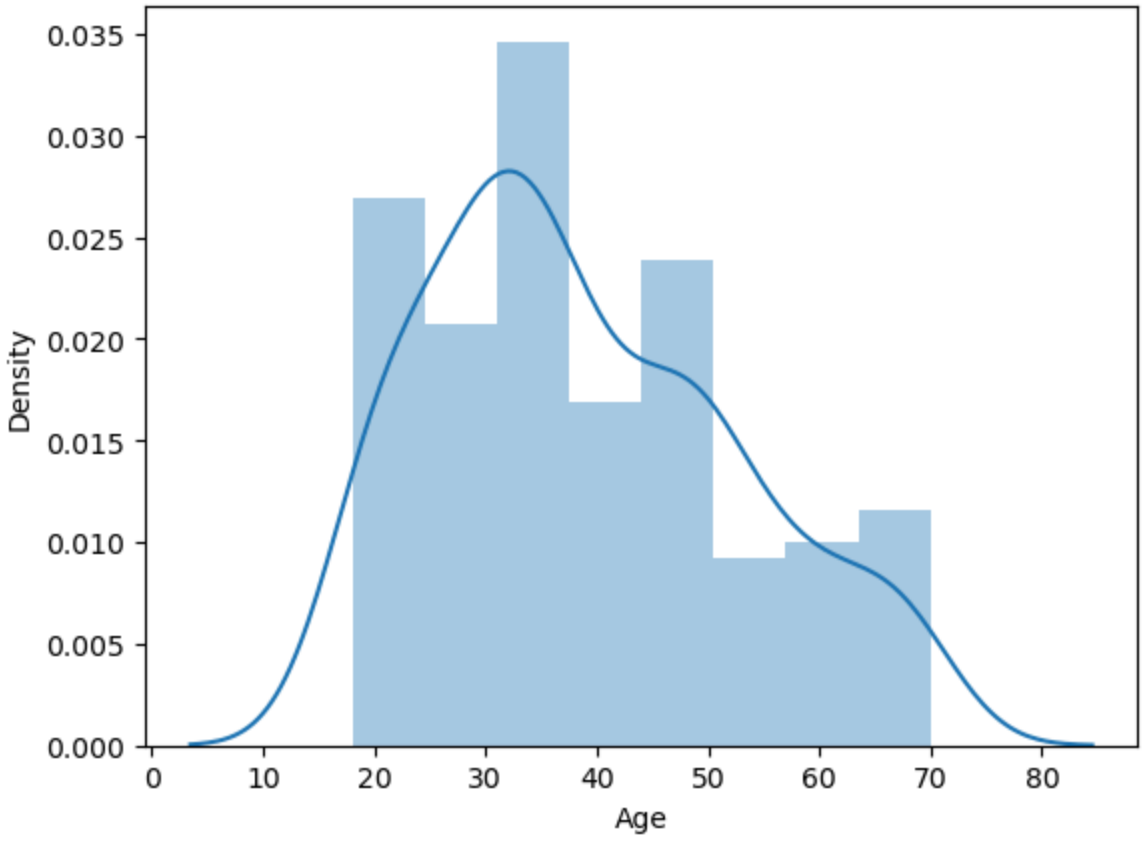
In [9]:
```
sns.distplot(df['Annual Income (k$)']);
```

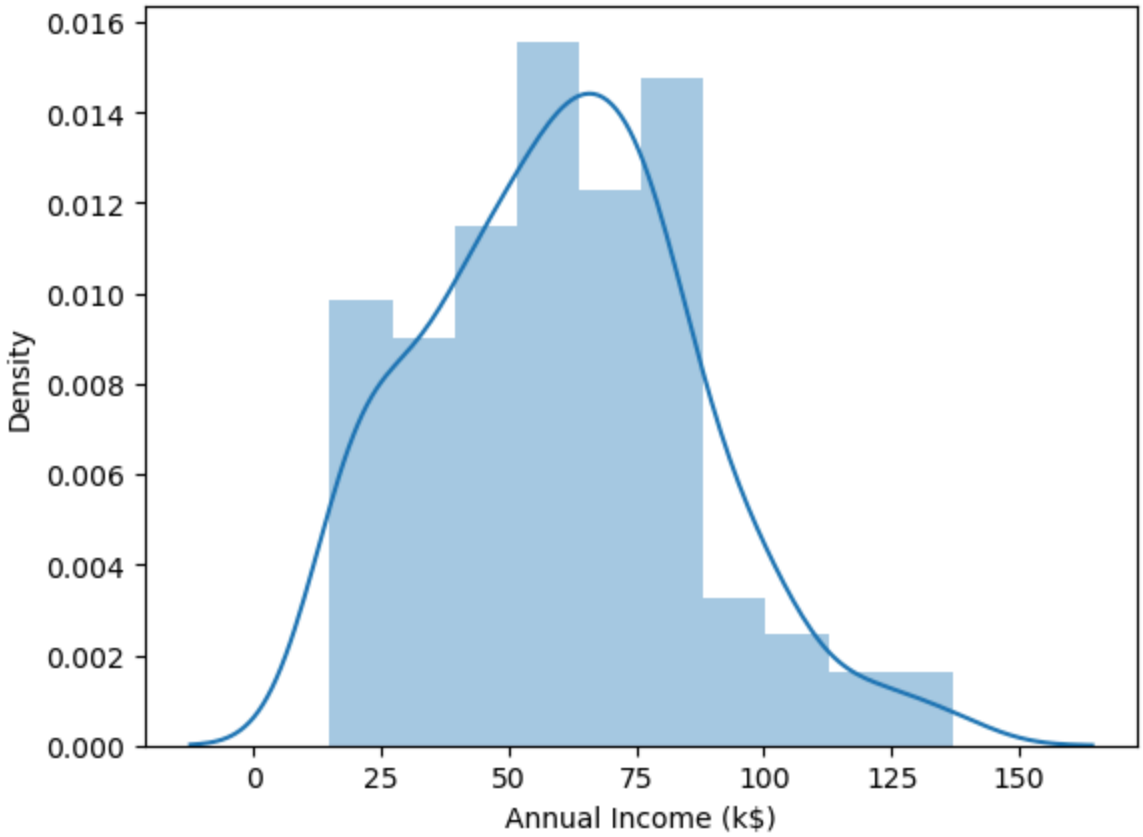


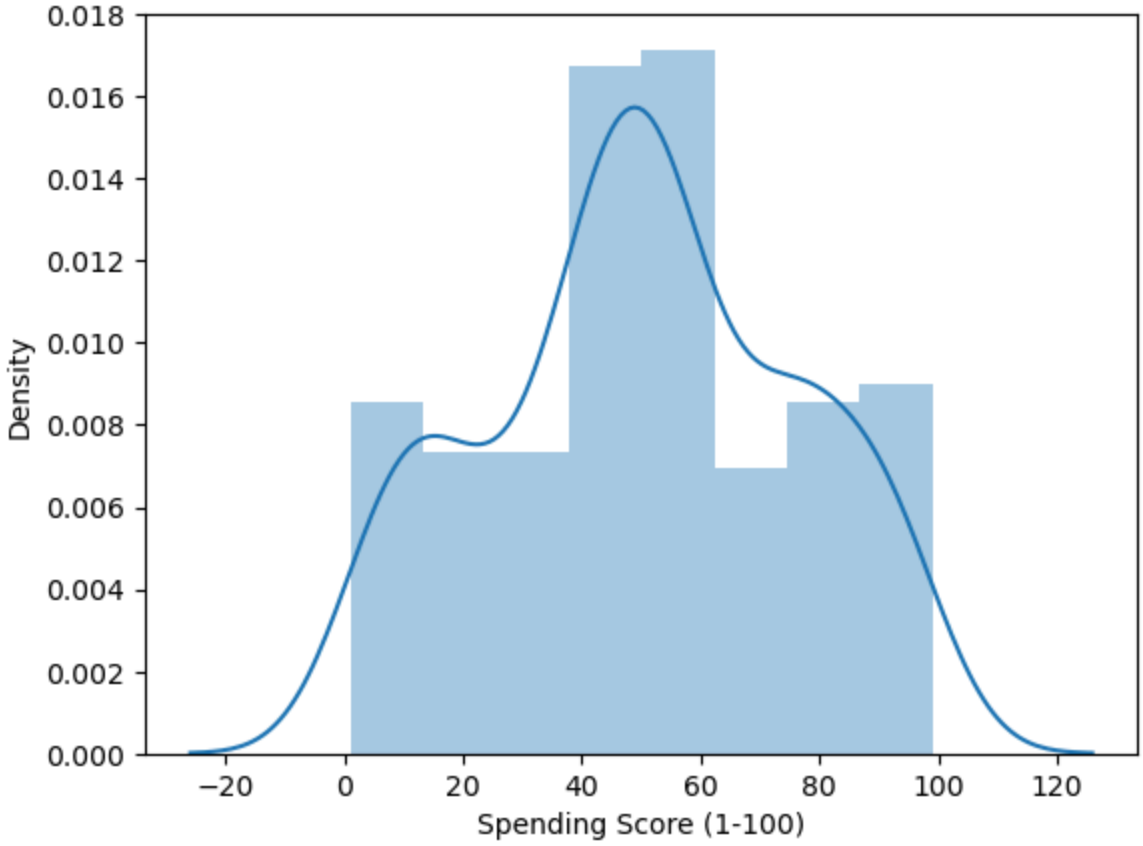
In [10]:
```
df.columns
```

Out[10]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)'],
      dtype='object')
```

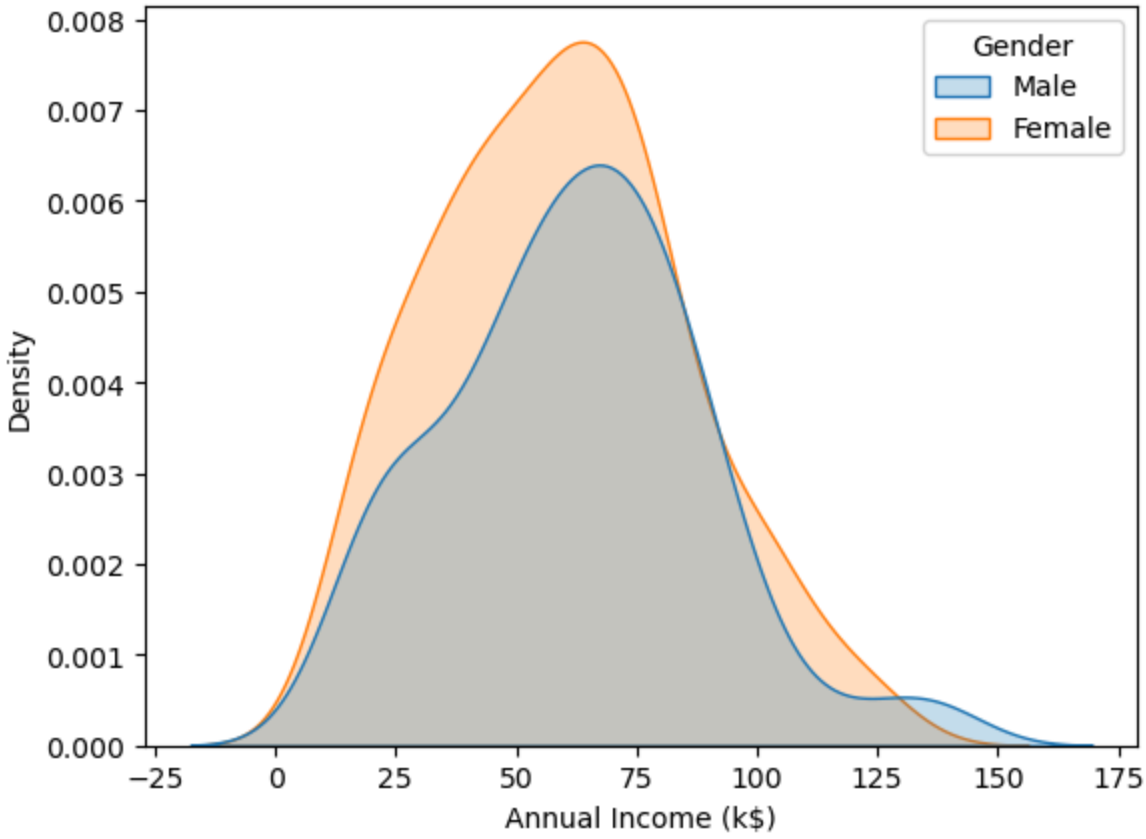
In [11]:
```
columns = ['Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.distplot(df[i])
```


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


0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0
25
50
75
100
125
150
Annual Income (k$)

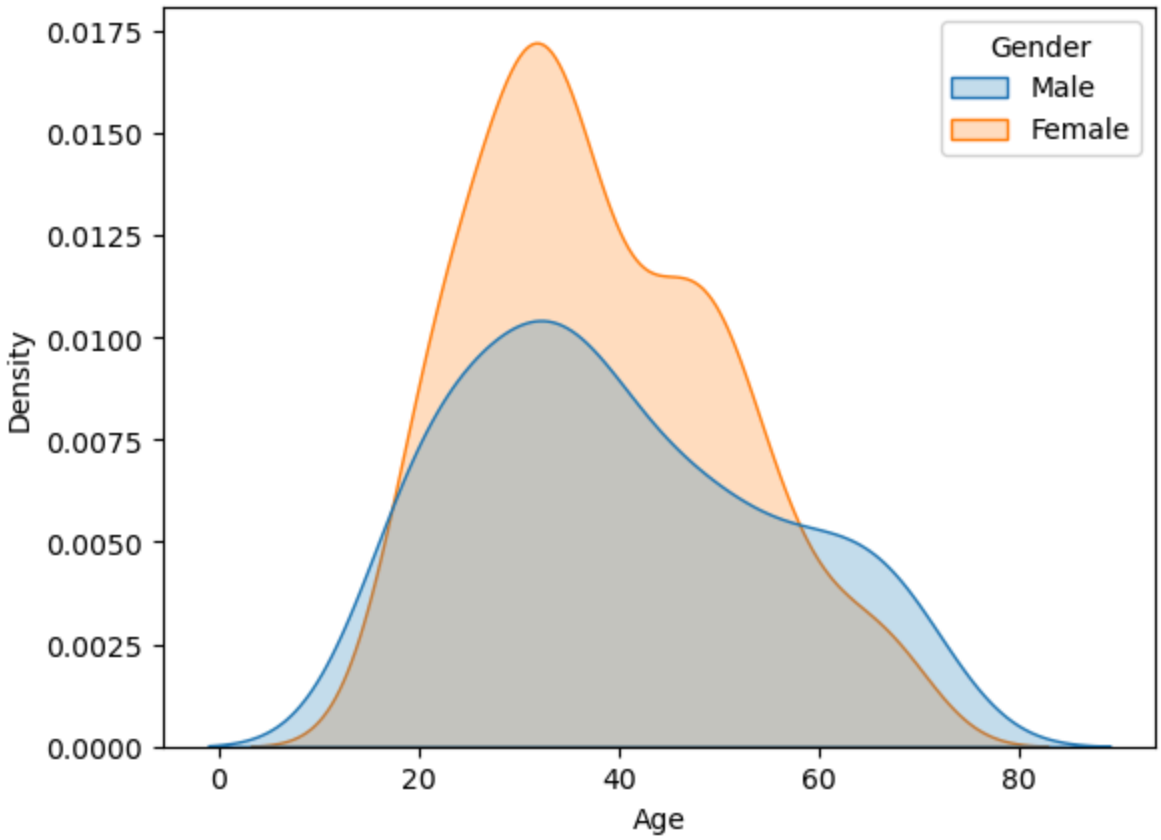


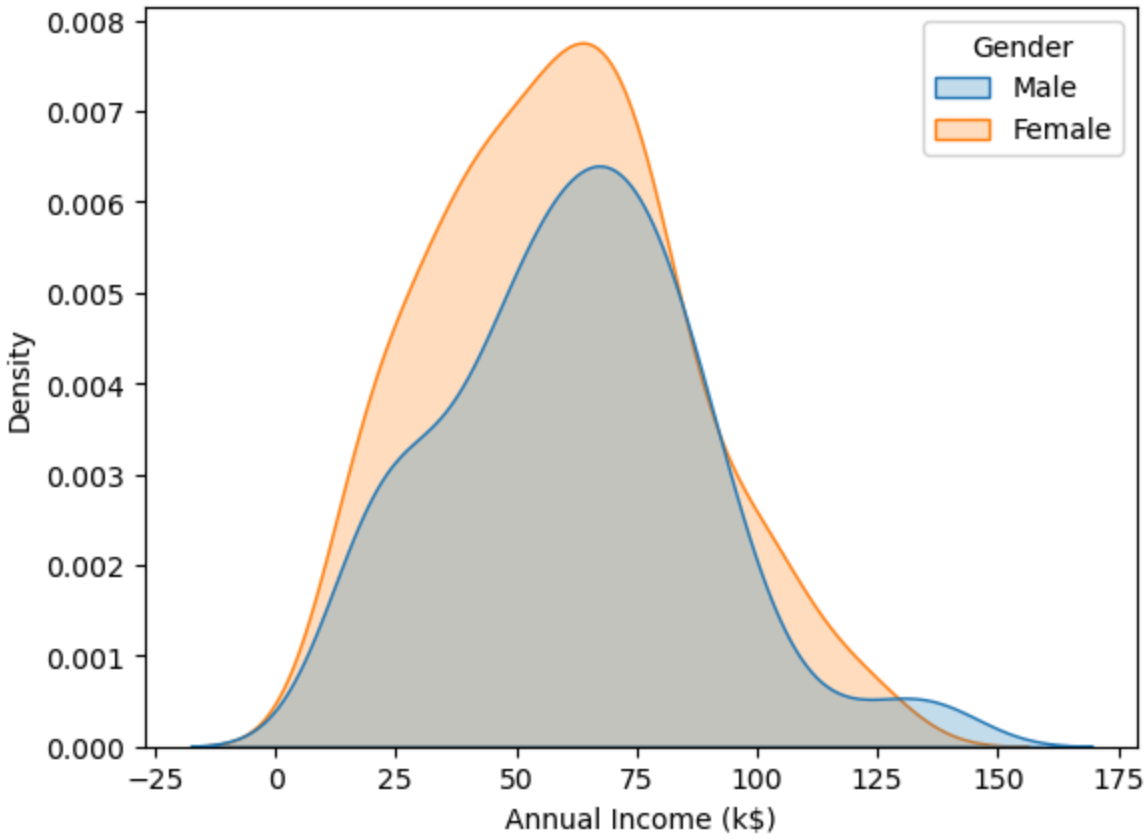
In [12]:
```
sns.kdeplot(df['Annual Income (k$)'],shade=True,hue=df['Gender']);
```

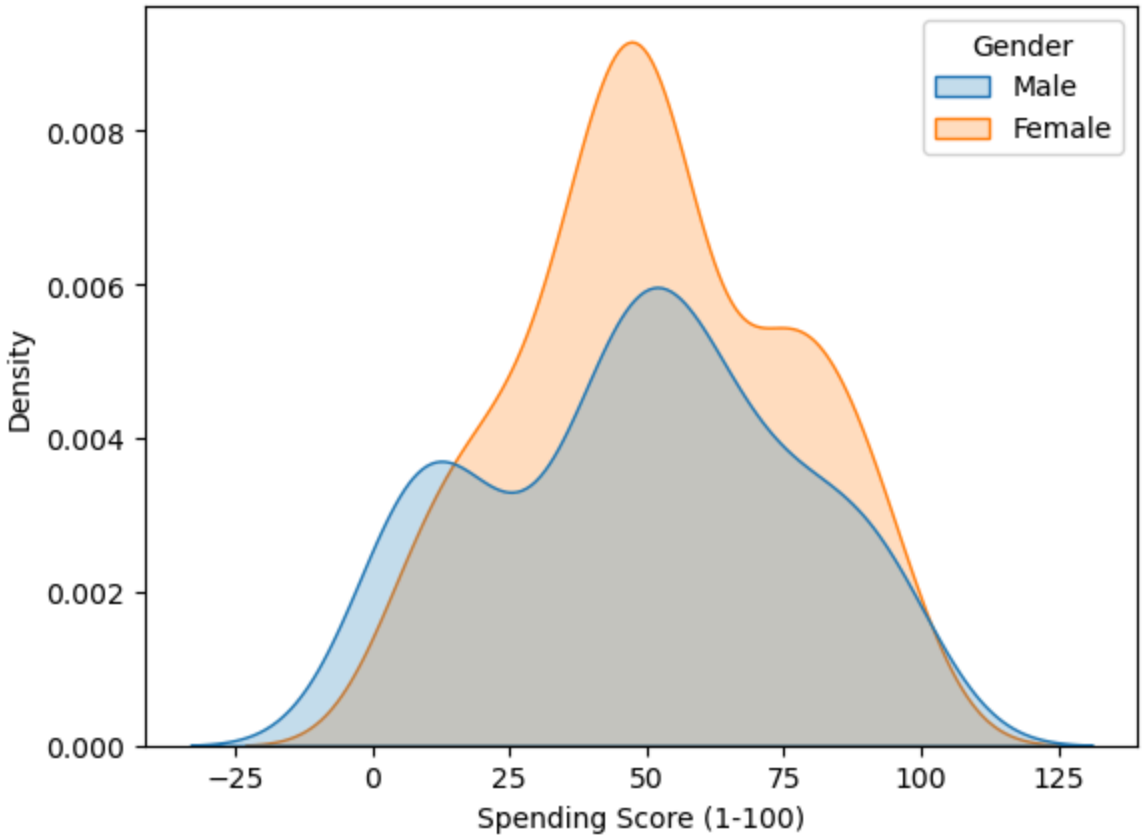


In [13]:
```
columns = ['Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in columns:
```

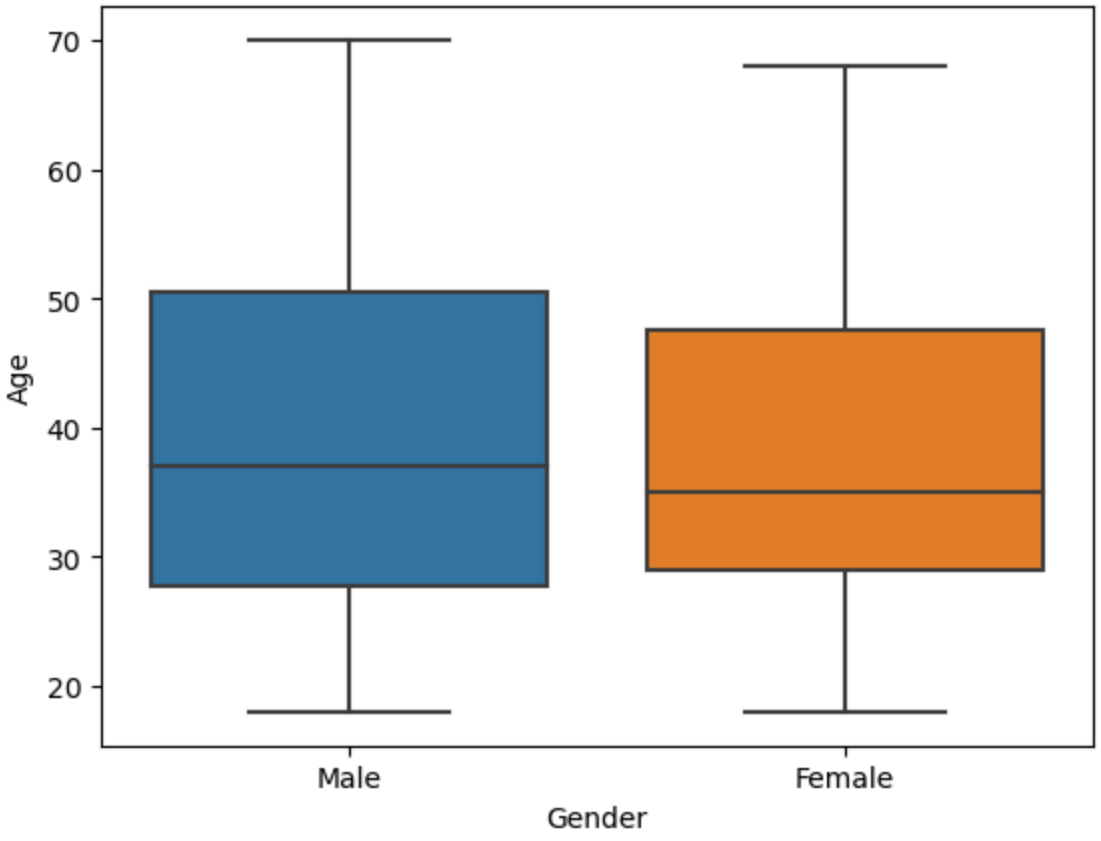
```
    plt.figure()
    sns.kdeplot(df[i],shade=True,hue=df['Gender'])
```

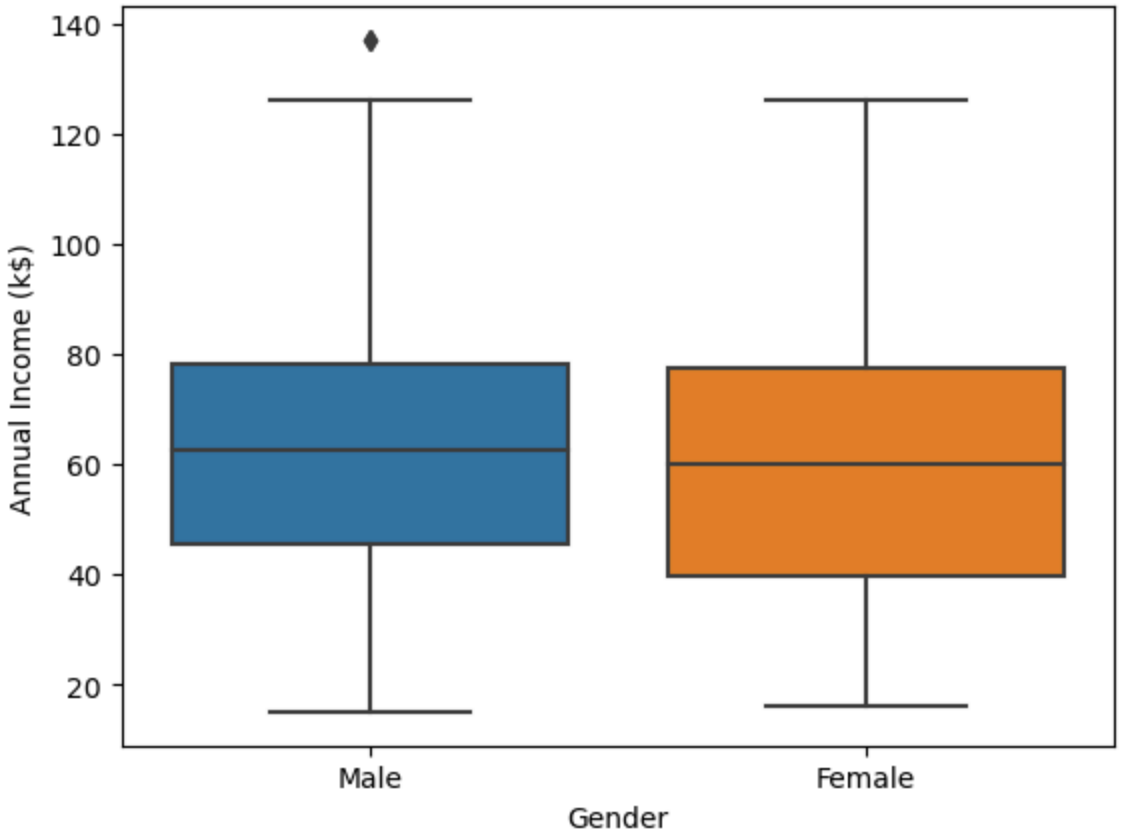


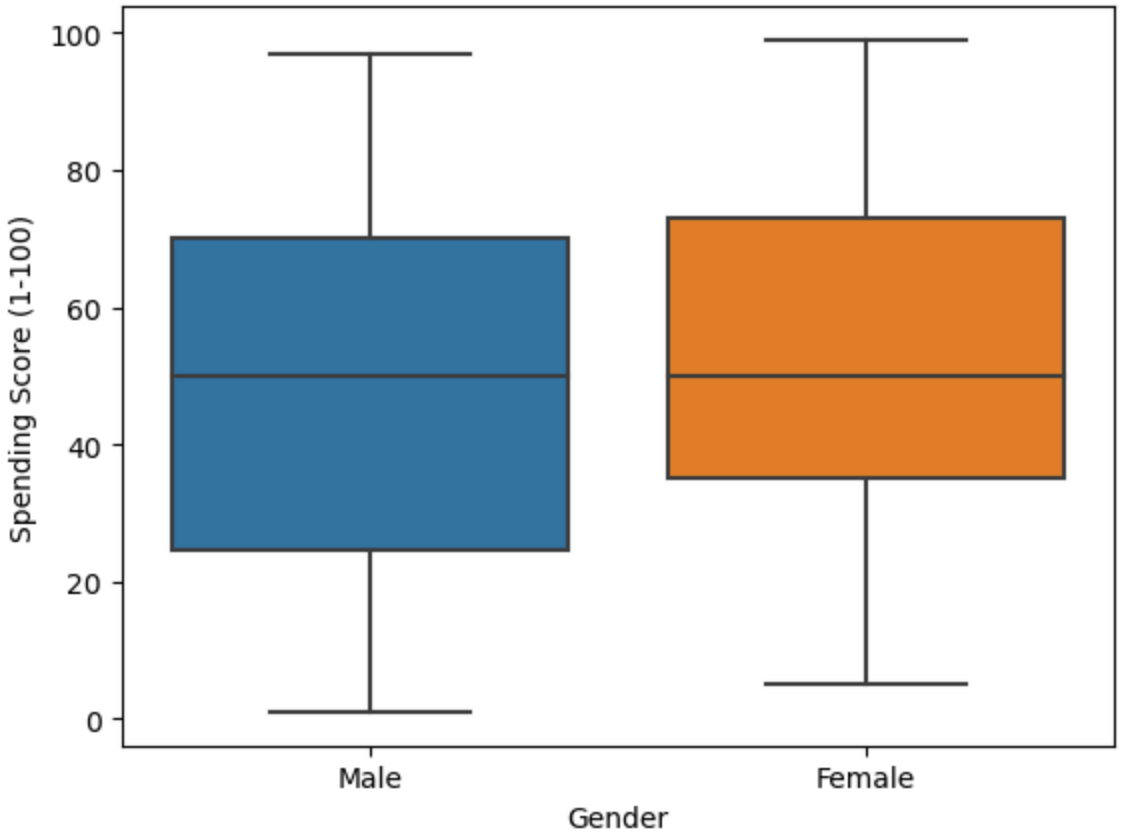




In [14]:
```
columns = ['Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.boxplot(data=df,x='Gender',y=df[i])
```





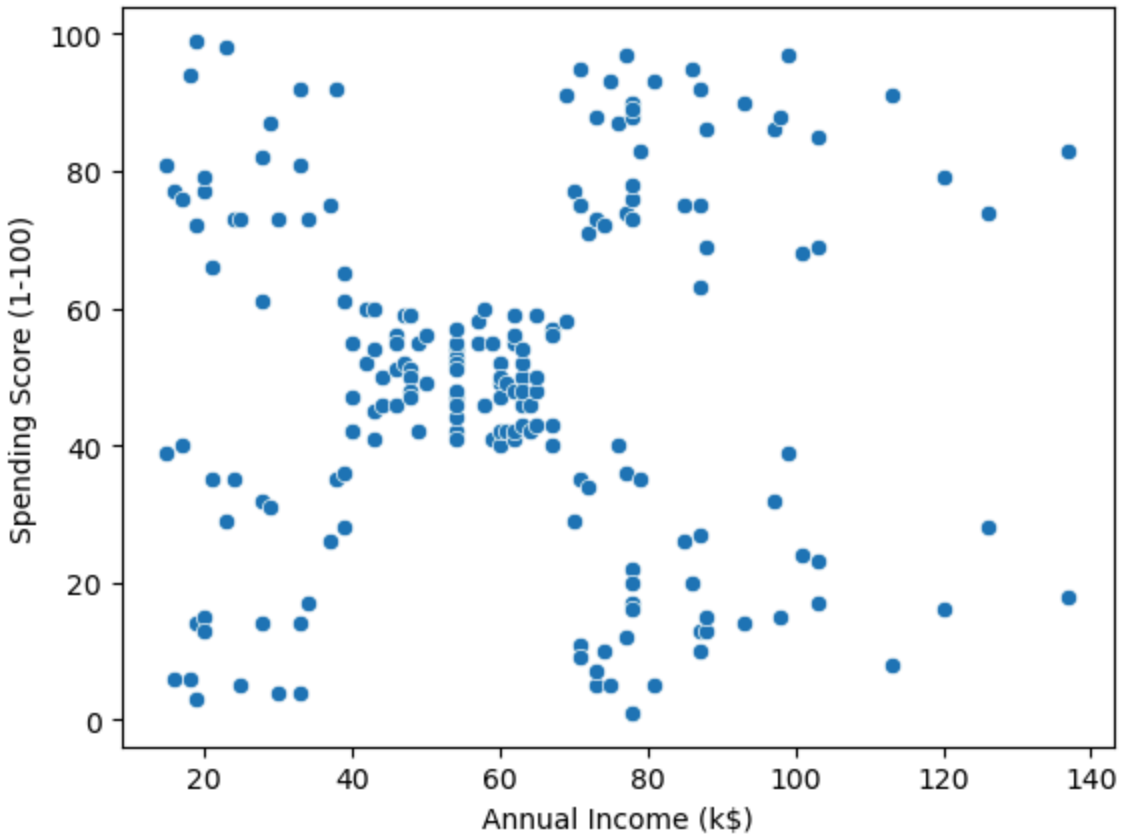


In [15]:
```
df['Gender'].value_counts(normalize=True)
```

Out[15]:
```
Female    0.56
Male      0.44
Name: Gender, dtype: float64
```

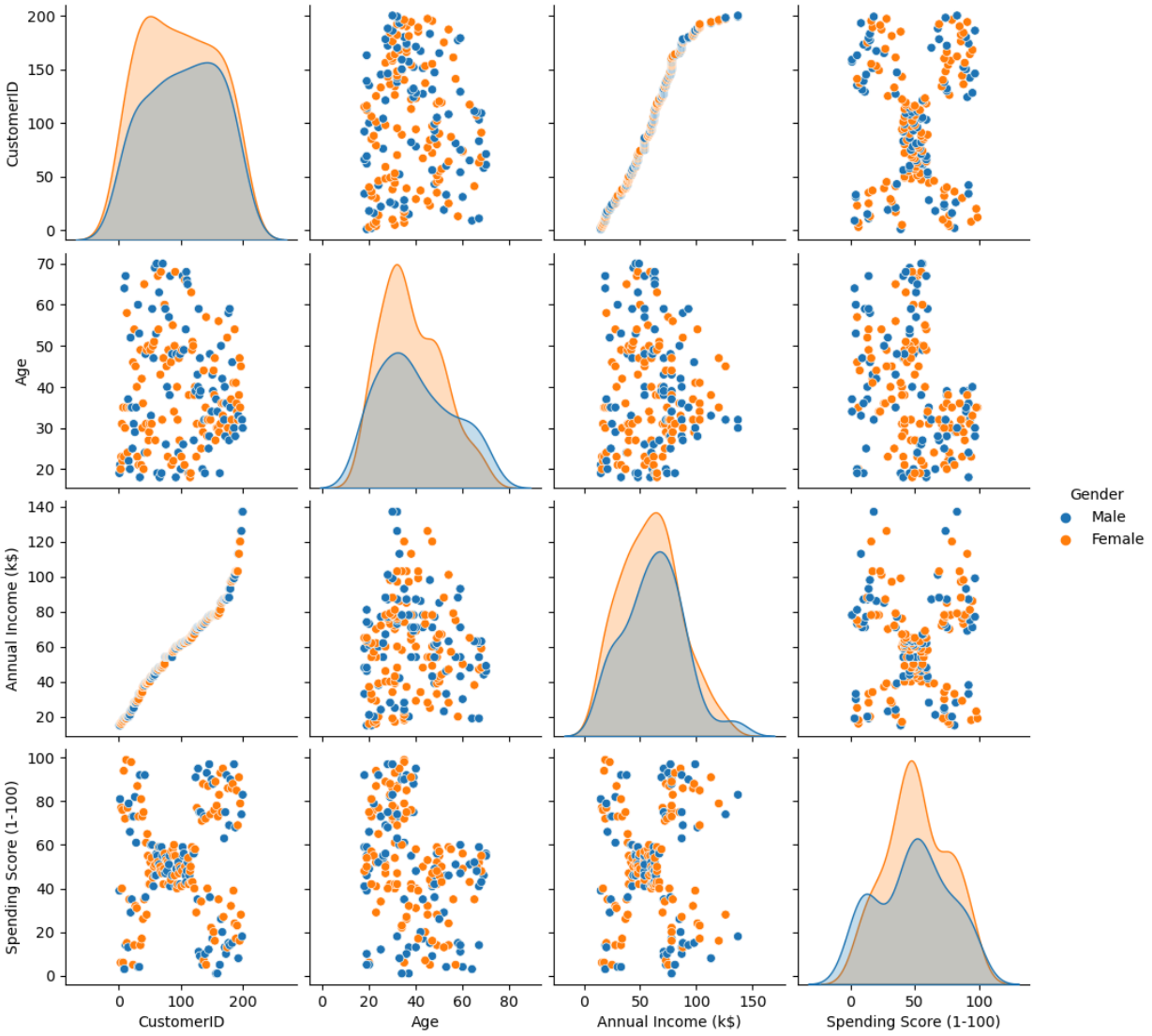
In [16]:
```
sns.scatterplot(data=df, x='Annual Income (k$)',y='Spending Score (1-100)' )
```

Out[16]:
```
<AxesSubplot:xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'>
```



In [17]:
```
#df=df.drop('CustomerID',axis=1)
sns.pairplot(df,hue='Gender')
```

Out[17]:
```
<seaborn.axisgrid.PairGrid at 0x7fdd5a50aa00>
```



In [18]:
```
df.groupby(['Gender'])['Age', 'Annual Income (k$)',
       'Spending Score (1-100)'].mean()
```

Out[18]:

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

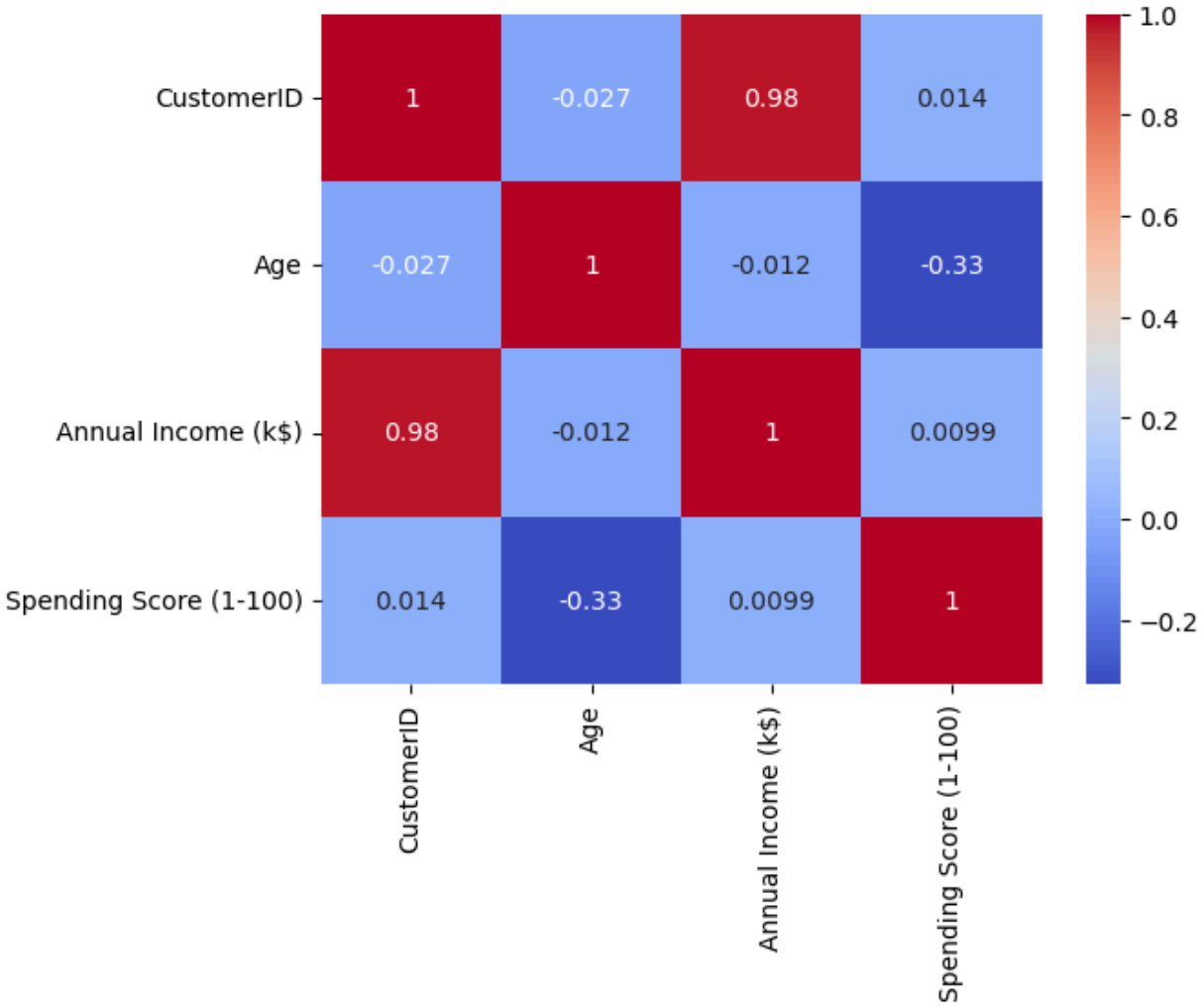
In [19]:
```
df.corr()
```

Out[19]:

	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 [20]:
```
sns.heatmap(df.corr(),annot=True,cmap='coolwarm')
```

Out[20]: <AxesSubplot:>



In [21]:
```
clustering1 = KMeans(n_clusters=3)
```

In [22]:
```
clustering1.fit(df[['Annual Income (k$)']])
```

Out[22]:
```
KMeans(n_clusters=3)
```

In [24]:
```
clustering1.labels_
```

Out[24]:
```
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, 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, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 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, 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, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2], dtype=int32)
```

In [25]:
```
df['Income Cluster'] = clustering1.labels_
df.head()
```

Out[25]:

	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 [26]:
```
df['Income Cluster'].value_counts()
```

Out[26]:
```
0    90
1    74
2    36
Name: Income Cluster, dtype: int64
```

In [27]:
```
clustering1.inertia_
```

Out[27]:
```
23517.330930930933
```

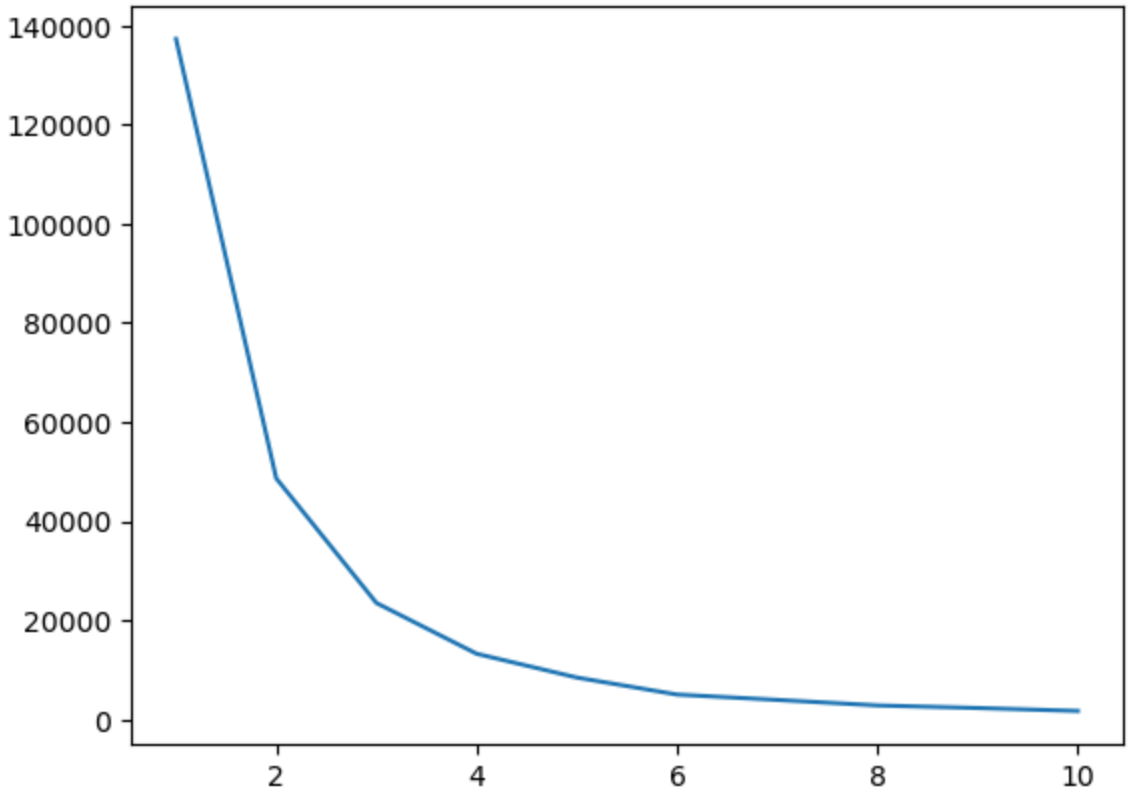
In [28]:
```
intertia_scores=[]
for i in range(1,11):
    kmeans=KMeans(n_clusters=i)
    kmeans.fit(df[['Annual Income (k$)']])
    intertia_scores.append(kmeans.inertia_)
```

In [29]:
```
intertia_scores
```

Out[29]:
```
[137277.28000000003,
 48660.88888888889,
 23517.330930930933,
 13278.112713472485,
 8481.496190476191,
 5050.9047619047615,
 3972.3214285714284,
 2857.441697191697,
 2335.8397186147185,
 1743.4772727272725]
```

In [30]:
```
plt.plot(range(1,11),intertia_scores)
```

Out[30]:
```
[<matplotlib.lines.Line2D at 0x7fdd6ae23820>]
```



In [31]:
```
df.columns
```

Out[31]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)', 'Income Cluster'],
      dtype='object')
```

In [32]:
```
df.groupby('Income Cluster')['Age', 'Annual Income (k$)',
       'Spending Score (1-100)'].mean()
```

Out[32]:

Income Cluster	Age	Annual Income (k$)	Spending Score (1-100)
0	38.722222	67.088889	50.000000
1	39.500000	33.486486	50.229730
2	37.833333	99.888889	50.638889

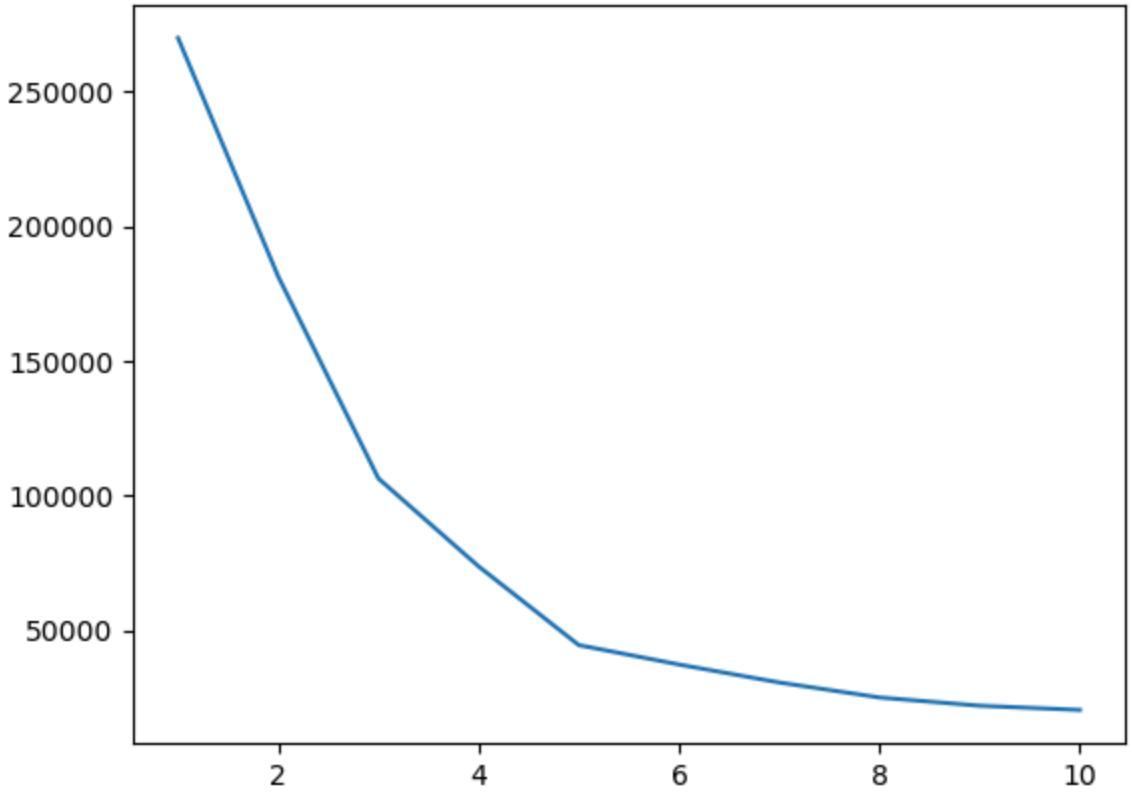
In [33]:
```
clustering2 = KMeans(n_clusters=5)
clustering2.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
df['Spending and Income Cluster'] =clustering2.labels_
df.head()
```

Out[33]:

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

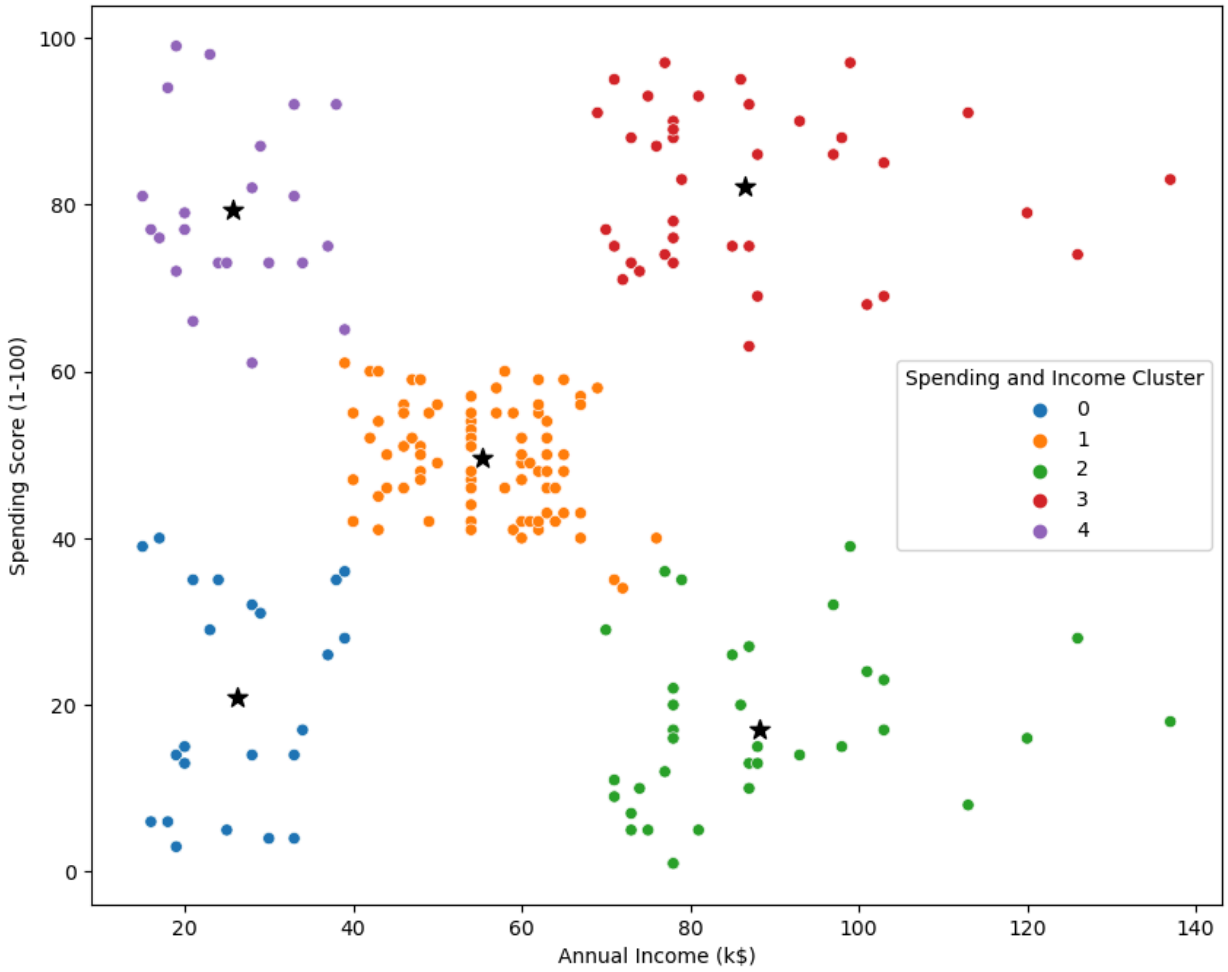
In [34]:
```
intertia_scores2=[]
for i in range(1,11):
    kmeans2=KMeans(n_clusters=i)
    kmeans2.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
    intertia_scores2.append(kmeans2.inertia_)
plt.plot(range(1,11),intertia_scores2)
```

Out[34]:
```
[<matplotlib.lines.Line2D at 0x7fdd5a5b8e20>]
```



In [35]:
```
centers =pd.DataFrame(clustering2.cluster_centers_)
centers.columns = ['x','y']
```

In [36]:
```
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)',hu
plt.savefig('clustering_bivaraiate.png')
```



In [37]:
```
pd.crosstab(df['Spending and Income Cluster'],df['Gender'],normalize='index')
```

Out[37]:

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

In [38]:
```
df.groupby('Spending and Income Cluster')['Age', 'Annual Income (k$)',
       'Spending Score (1-100)'].mean()
```

Out[38]:

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

In [39]:
```
#mulivariate clustering
from sklearn.preprocessing import StandardScaler
```

In [40]:
```
scale = StandardScaler()
```

In [41]:
```
df.head()
```

Out[41]:

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

In [42]:
```
dff = pd.get_dummies(df,drop_first=True)
dff.head()
```

Out[42]:

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

In [43]:
```
dff.columns
```

Out[43]:
```
Index(['CustomerID', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)',
       'Income Cluster', 'Spending and Income Cluster', 'Gender_Male'],
      dtype='object')
```

In [44]:
```
dff = dff[['Age', 'Annual Income (k$)', 'Spending Score (1-100)','Gender_Male']
dff.head()
```

Out[44]:

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

In [45]:
```
dff = scale.fit_transform(dff)
```

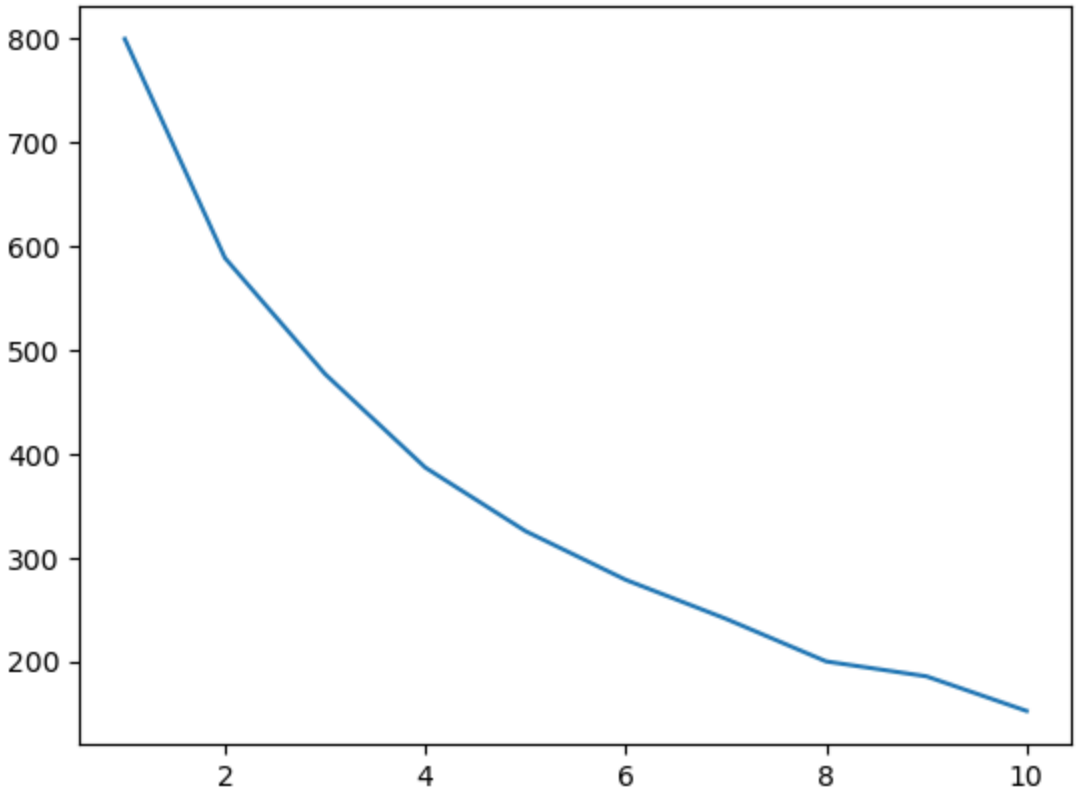
In [46]:
```
dff = pd.DataFrame(scale.fit_transform(dff))
dff.head()
```

Out[46]:

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

In [47]:
```
intertia_scores3=[]
for i in range(1,11):
    kmeans3=KMeans(n_clusters=i)
    kmeans3.fit(dff)
    intertia_scores3.append(kmeans3.inertia_)
plt.plot(range(1,11),intertia_scores3)
```

Out[47]:
```
[<matplotlib.lines.Line2D at 0x7fdd78b0aca0>]
```



In [48]:
```
df
```

Out[48]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	Income Cluster	Spending and Income Cluster
0	1	Male	19	15	39	1	0
1	2	Male	21	15	81	1	4
2	3	Female	20	16	6	1	0
3	4	Female	23	16	77	1	4
4	5	Female	31	17	40	1	0
...	...	...	...	...	...	...	...
195	196	Female	35	120	79	2	3
196	197	Female	45	126	28	2	2
197	198	Male	32	126	74	2	3
198	199	Male	32	137	18	2	2
199	200	Male	30	137	83	2	3

200 rows × 7 columns

In [49]:
```
df.to_csv('Clustering.csv')
```

In [ ]:
```

```