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/pankt/OneDrive/Documents/Projects/Data_Analytics/Datasets/
```

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

Univariate Analysis

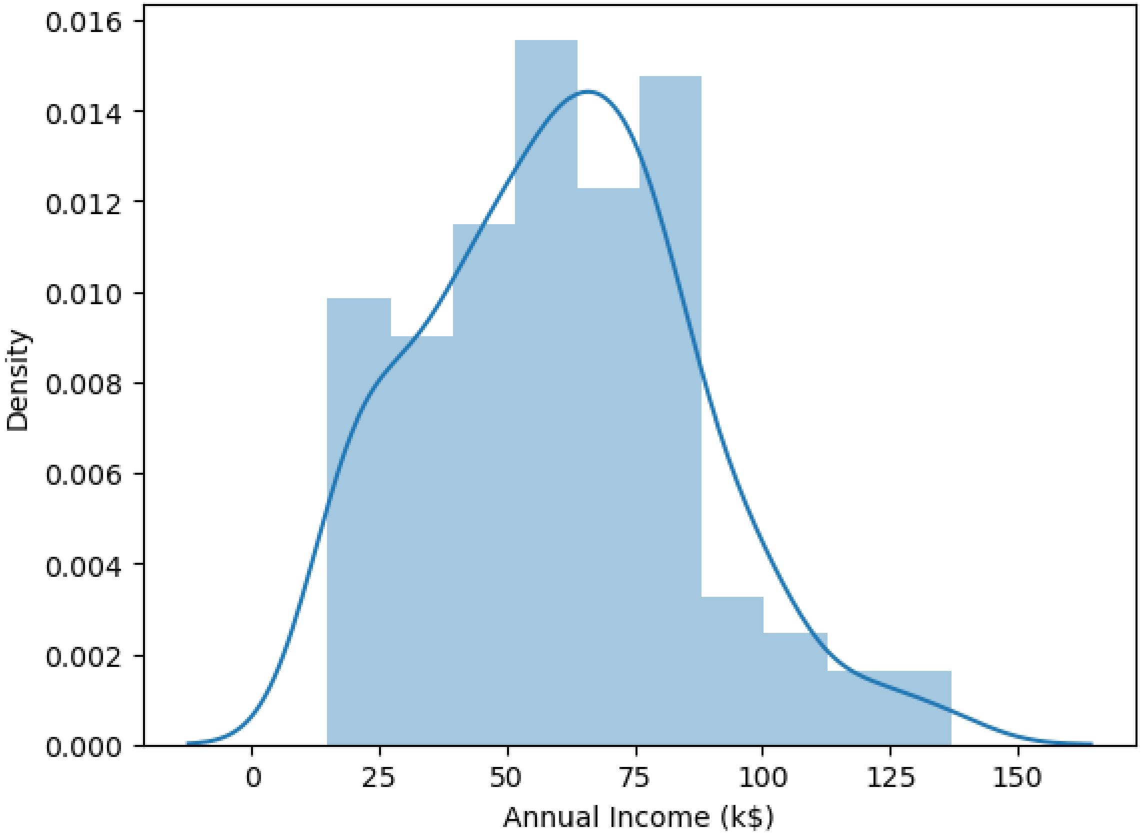
In [4]:
```
df.describe()
```

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$)'])
```

Out[5]:
```
<AxesSubplot:xlabel='Annual Income (k$)', ylabel='Density'>
```

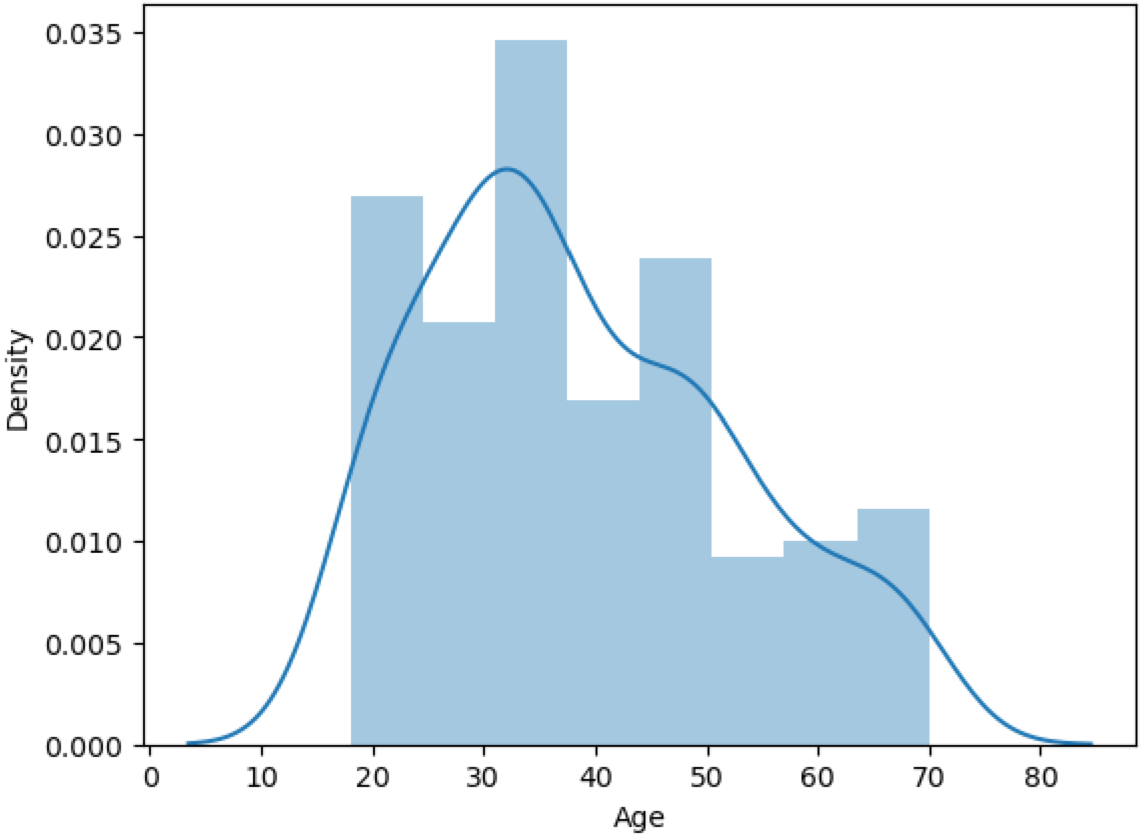


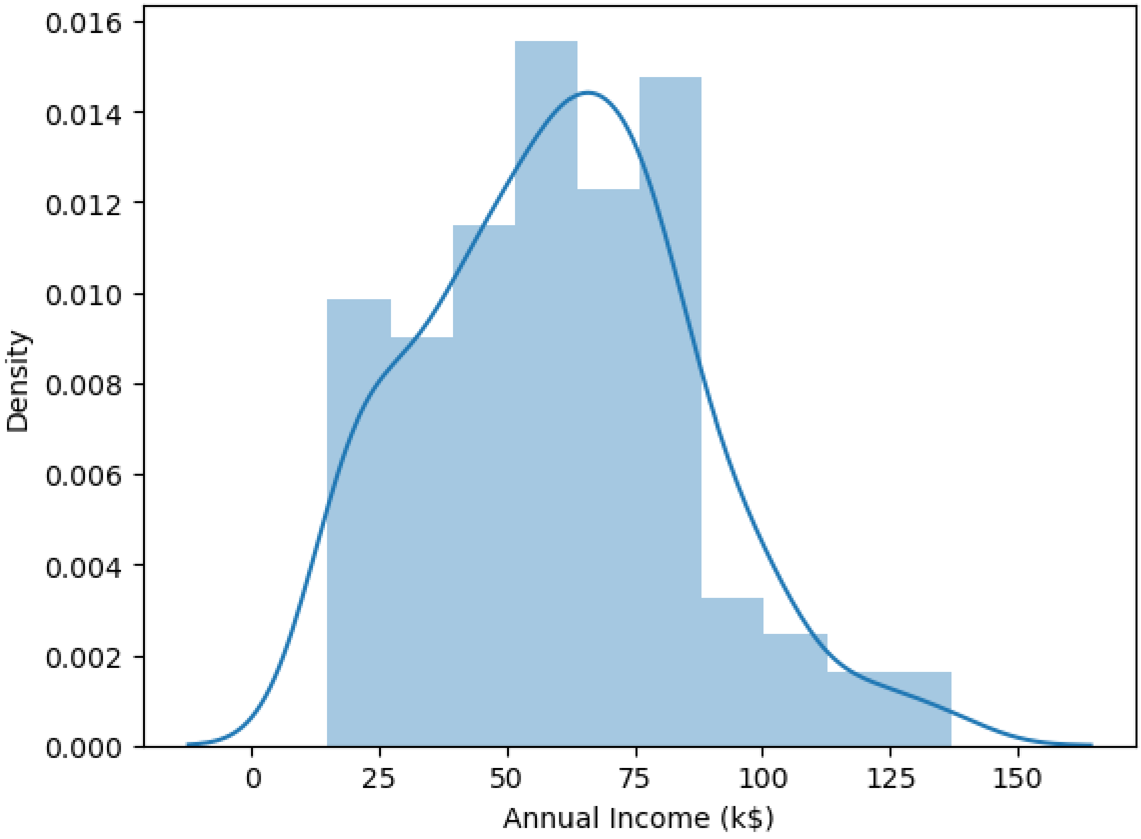
In [6]:
```
df.columns
```

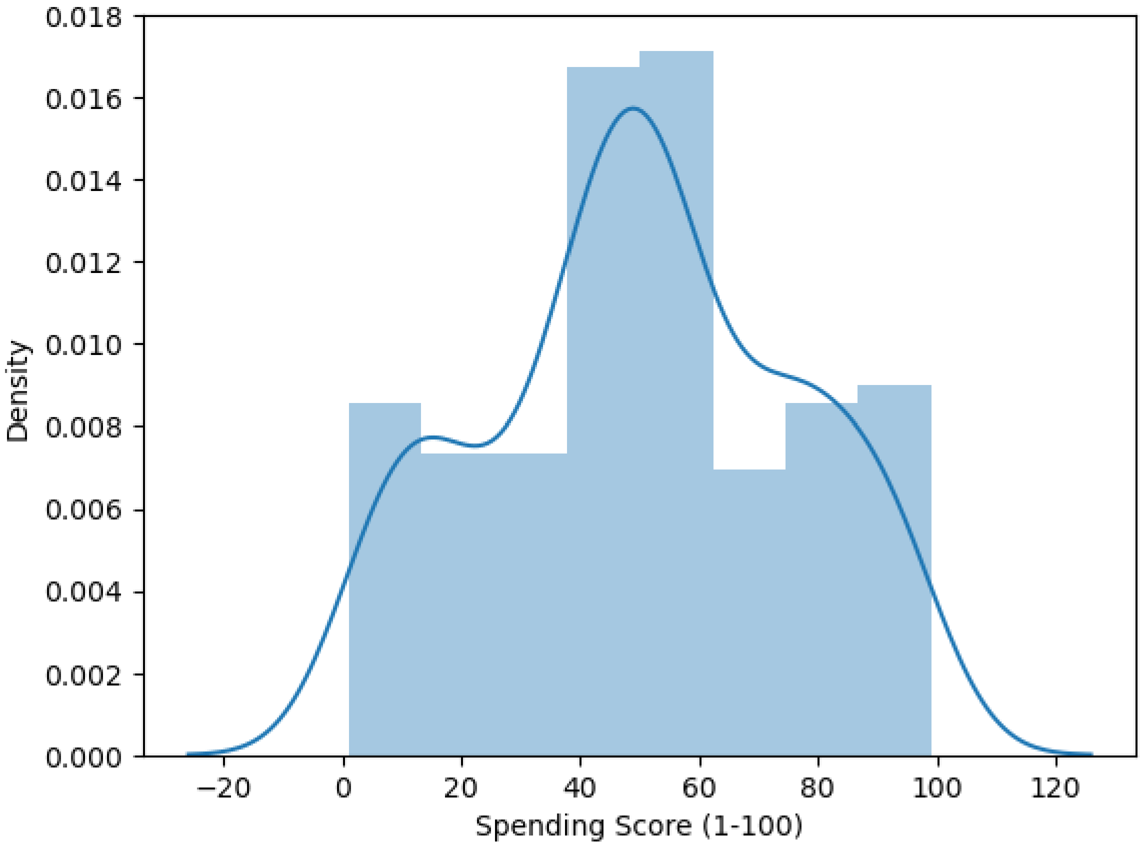
Out[6]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)'],
      dtype='object')
```

To plot multiple columns

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

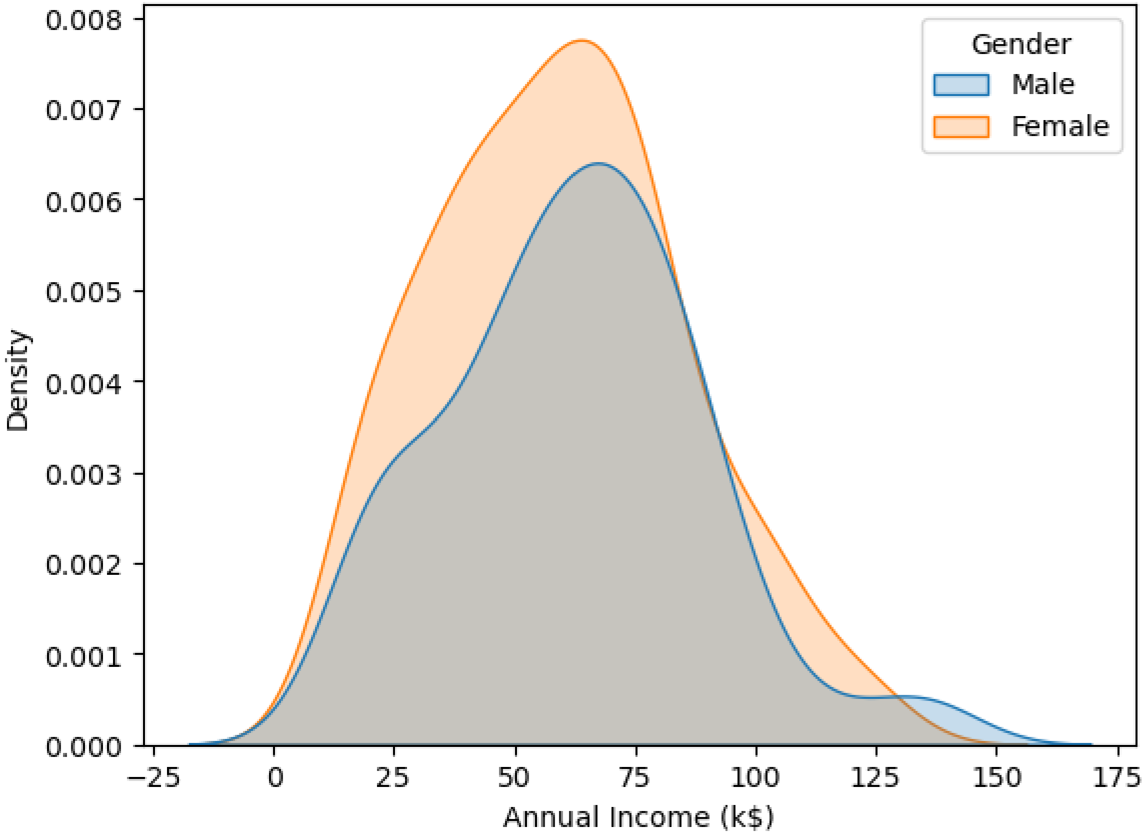

0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
10
20
30
40
50
60
70
80
Age


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$)



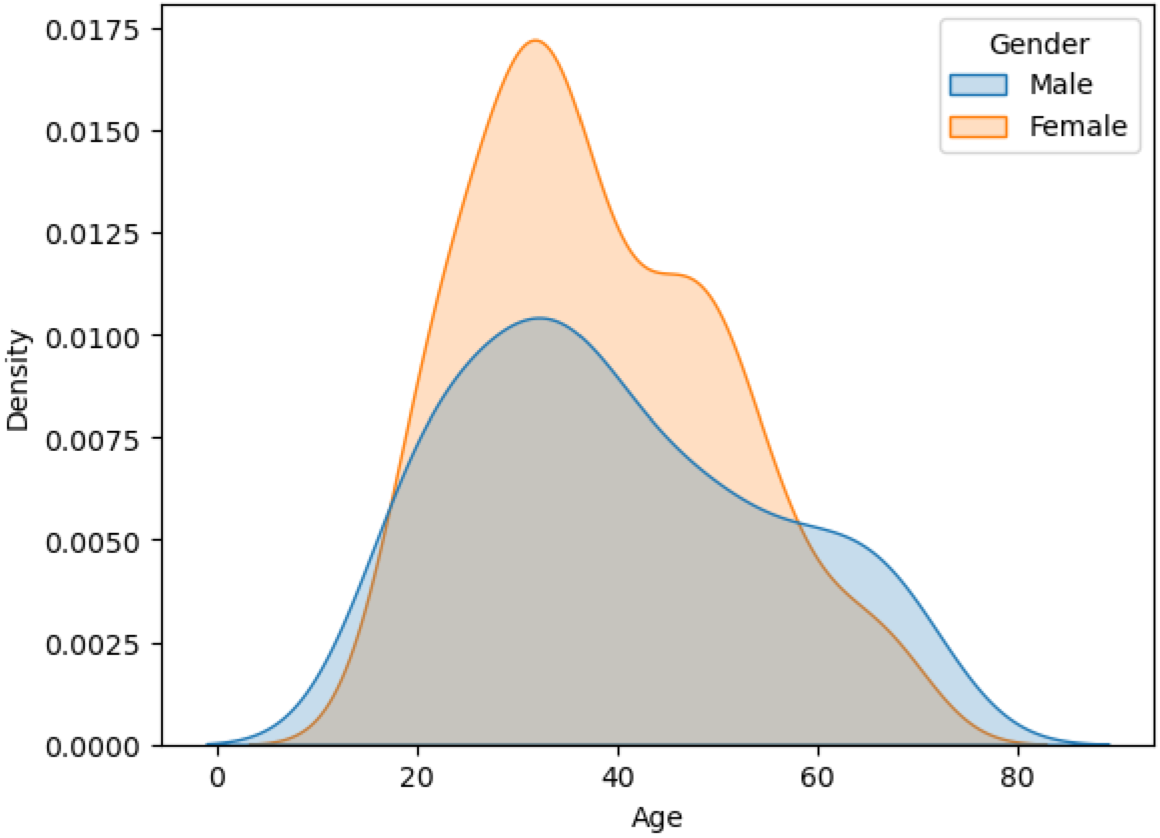
Ploting the variables by gender

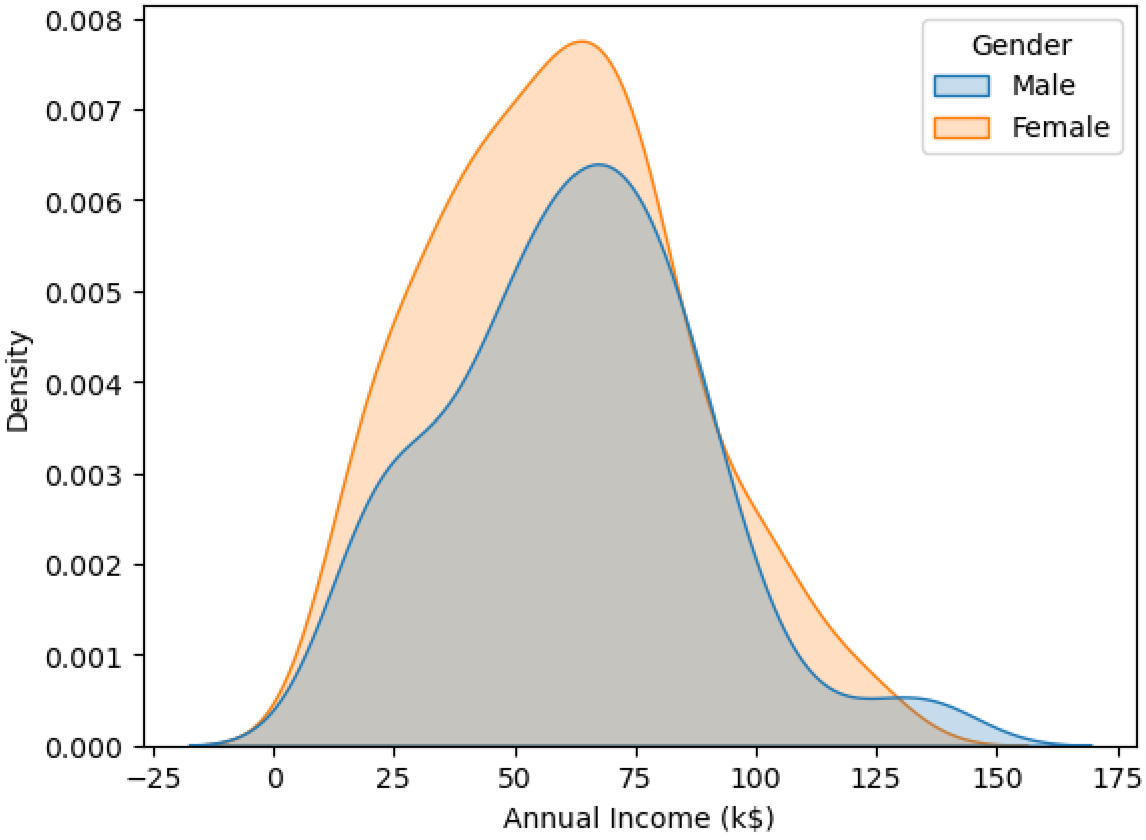
In [10]:
```
sns.kdeplot(df['Annual Income (k$)'], shade=True, hue=df['Gender']);
```

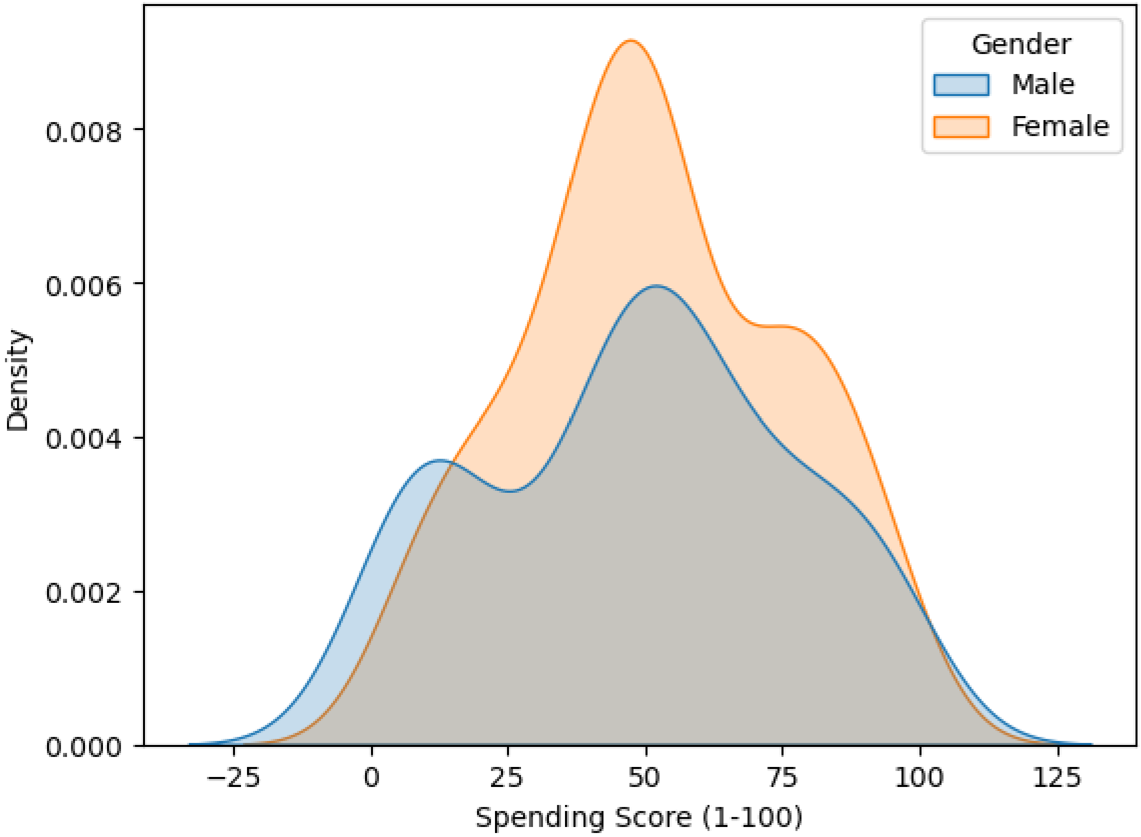


In [11]:

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

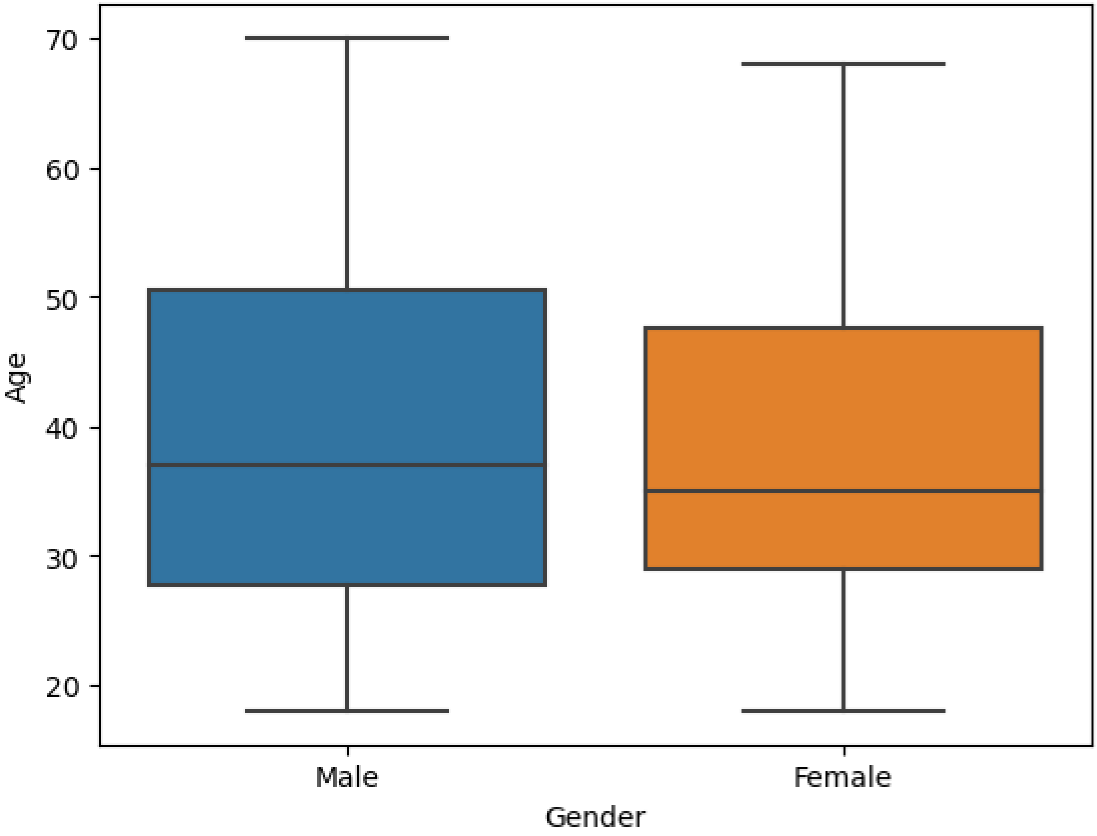


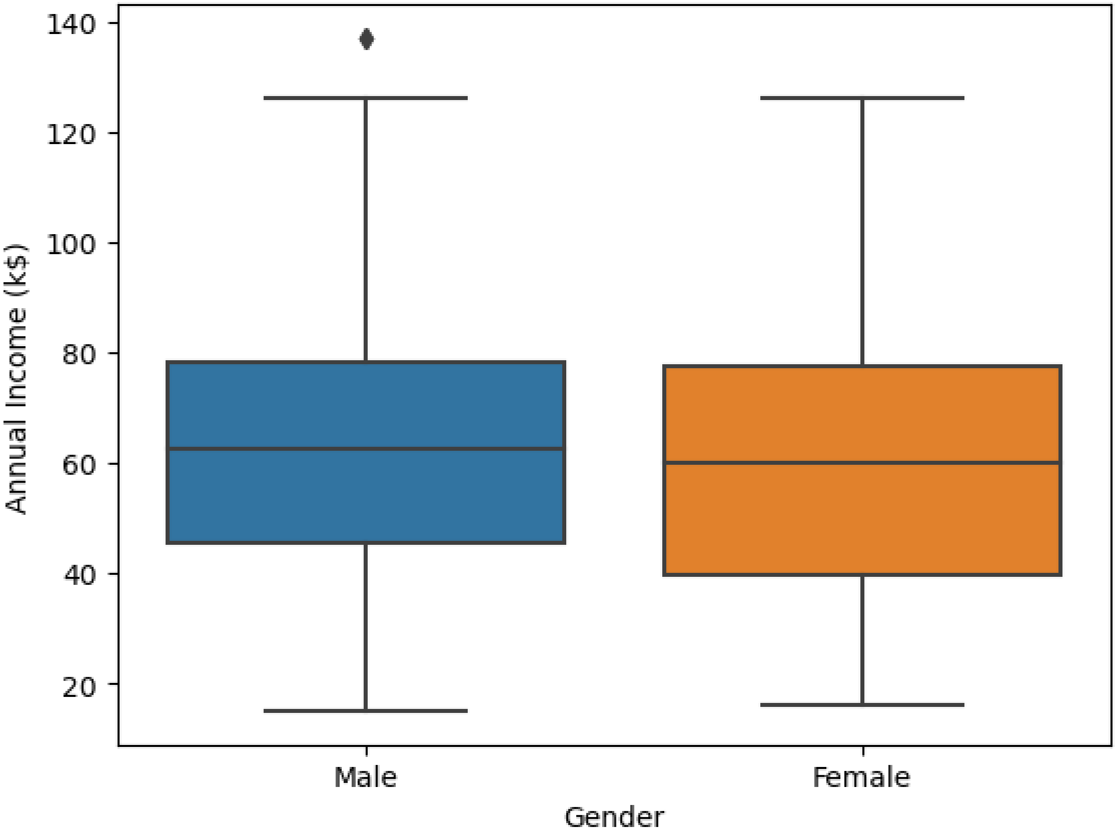


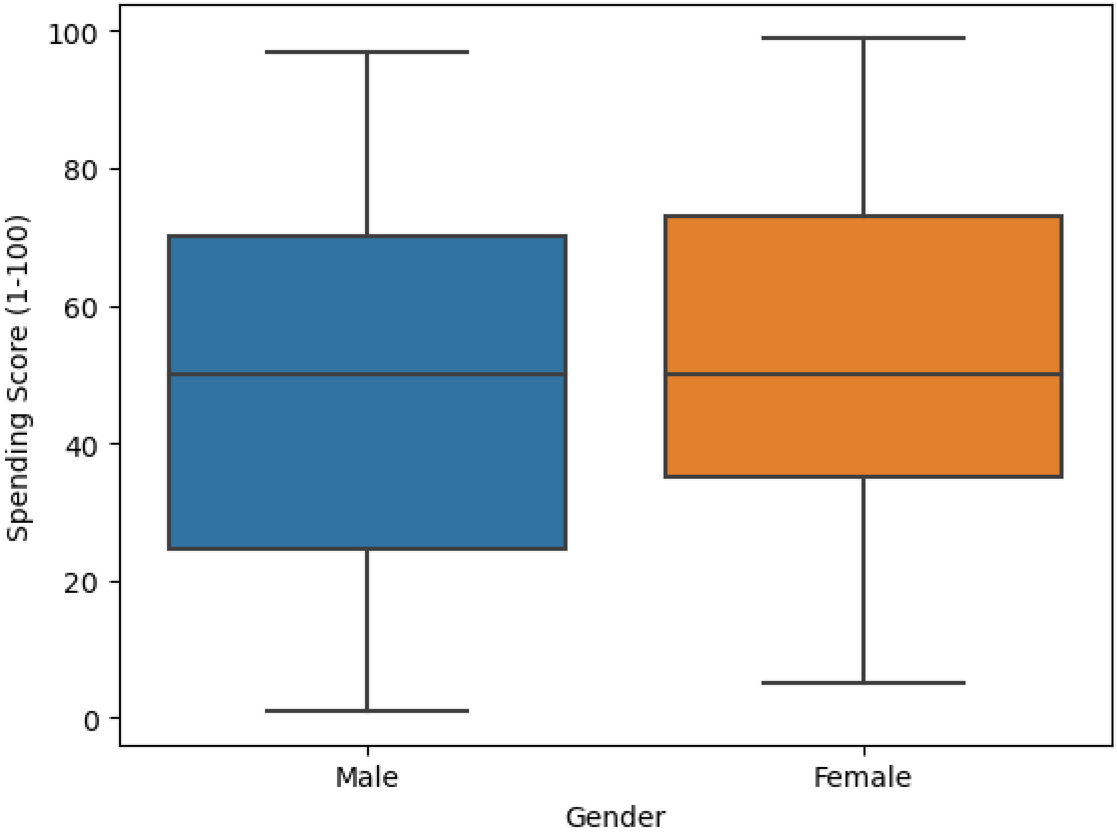


In [12]:

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


70
60
50
40
30
20
Age
Male
Female
Gender


140
120
100
80
60
40
20
Annual Income (k$)
Male
Female
Gender



In [14]:
```
df['Gender'].value_counts(normalize=True)
```

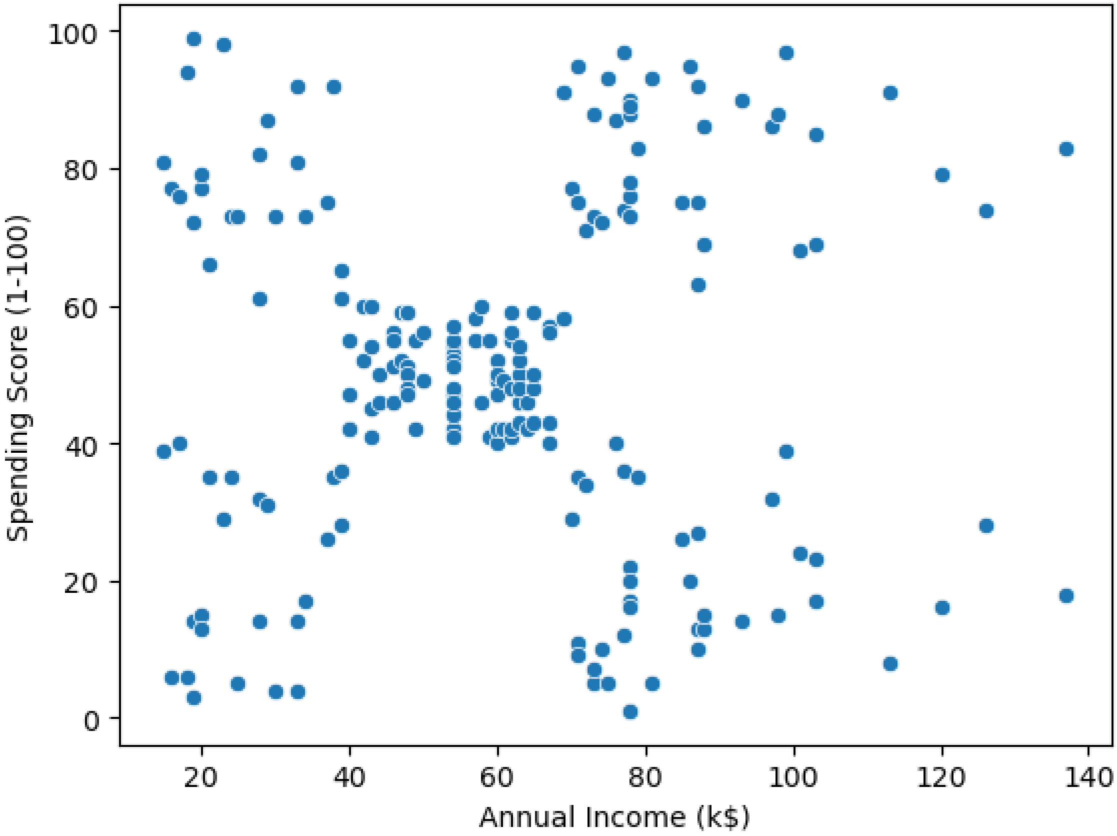
Out[14]:
```
Female    0.56
Male      0.44
Name: Gender, dtype: float64
```

In [ ]:

Bivariate Analysis

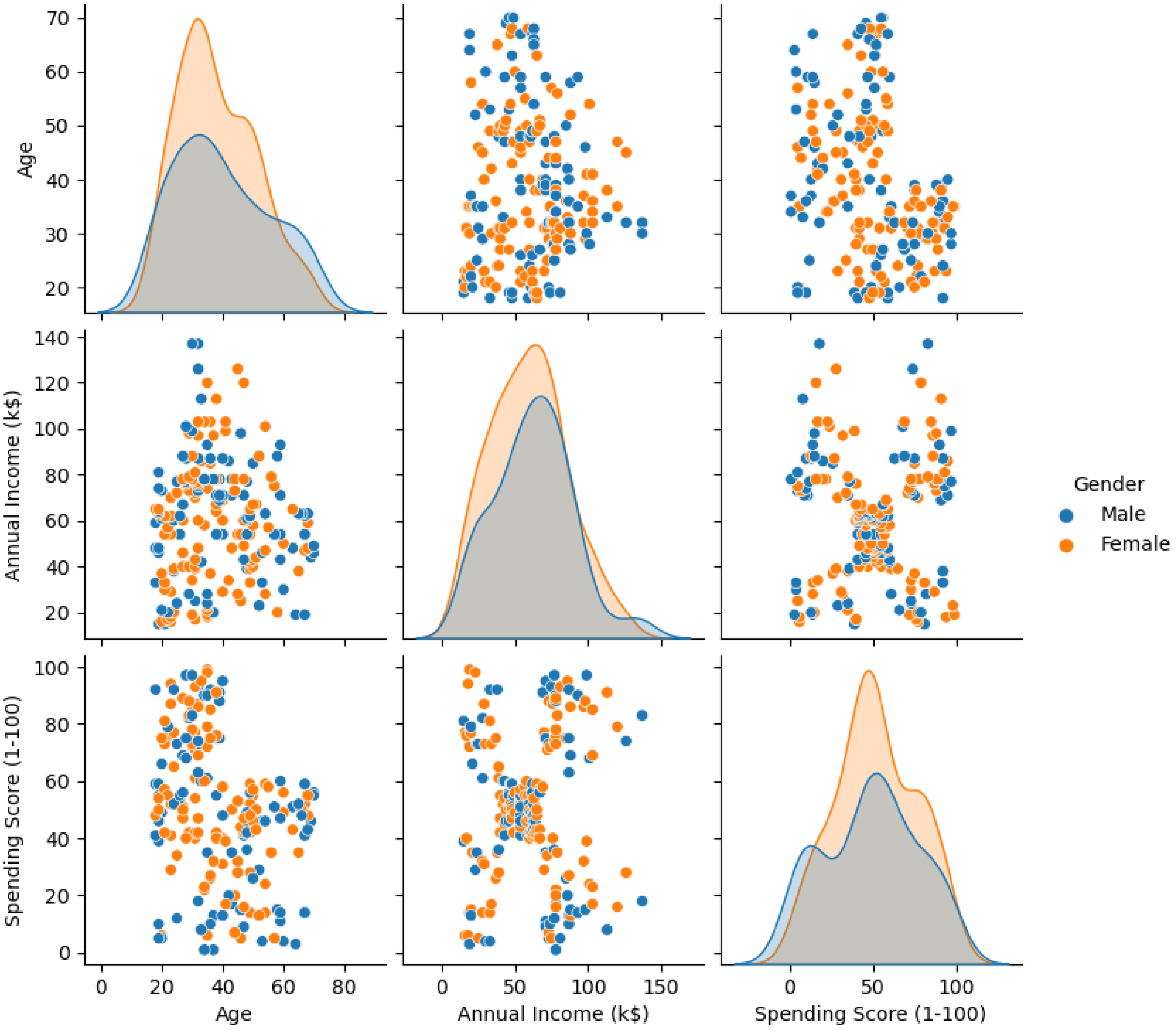
In [15]:
```
sns.scatterplot(data=df, x='Annual Income (k$)', y = 'Spending Score (1-100)')
```

Out[15]:
```
<AxesSubplot:xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'>
```



In [21]:
```
# df=df.drop('CustomerID', axis=1)
sns.pairplot(df, hue='Gender')
```

Out[21]:
```
<seaborn.axisgrid.PairGrid at 0x253d5d70d60>
```



In [23]:

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

Out[23]:

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

In [24]:

```
df.corr()
```

Out[24]:

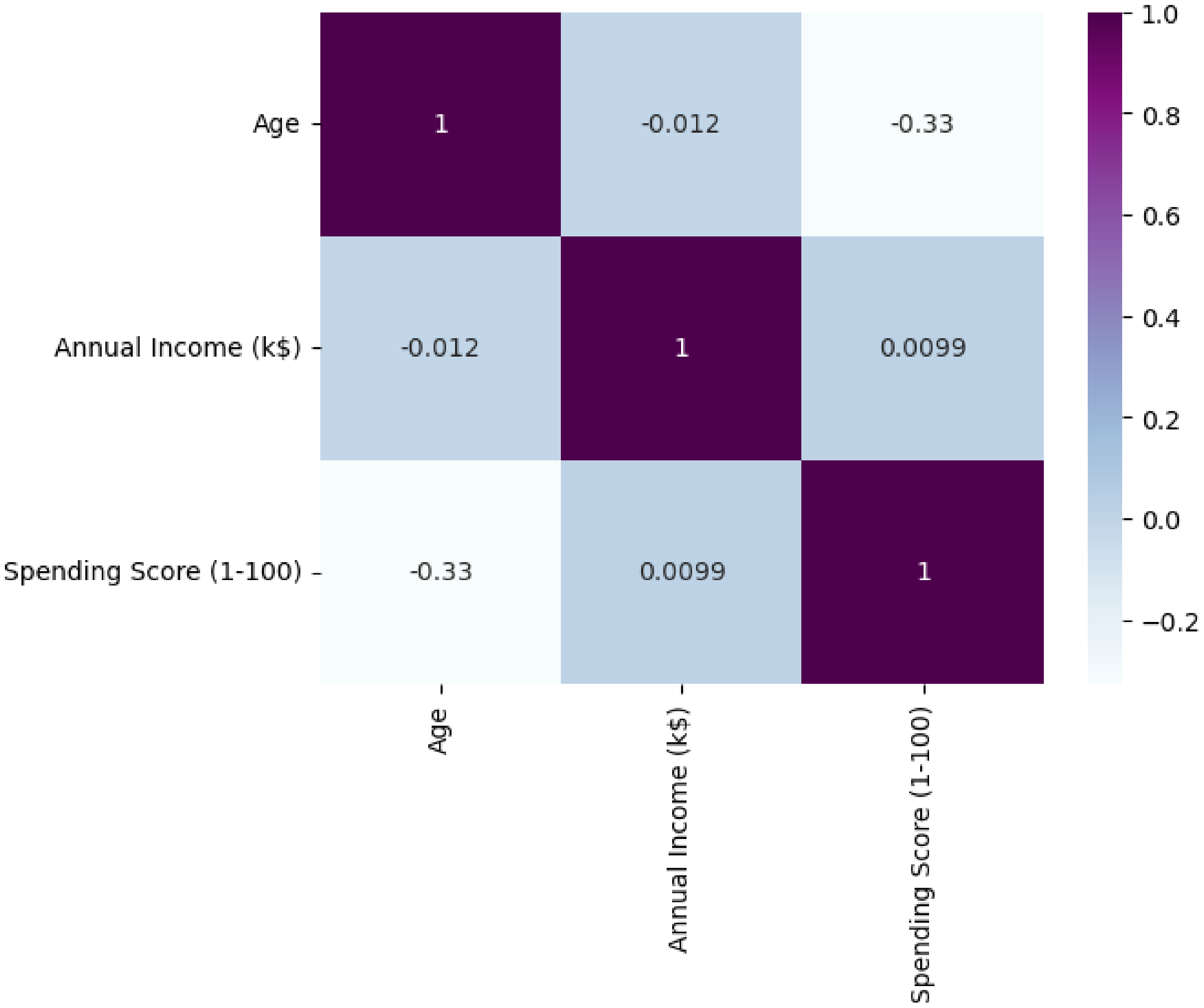
	Age	Annual Income (k$)	Spending Score (1-100)
Age	1.000000	-0.012398	-0.327227
Annual Income (k$)	-0.012398	1.000000	0.009903
Spending Score (1-100)	-0.327227	0.009903	1.000000

In [30]:

```
sns.heatmap(df.corr(),annot=True, cmap='BuPu')
```

Out[30]:

```
<AxesSubplot:>
```



Clustering - Univariate, Bivariate, Multivariate

In [52]:

```
clustering1 = KMeans(n_clusters=3)
```

In [53]:

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

Out[53]:

```
KMeans(n_clusters=3)
```

In [54]:

```
clustering1.labels_
```

Out[54]:

```
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, 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])
```

In [55]:

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

Out[55]:

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

In [56]:

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

Out[56]:

```
0    90
1    74
2    36
Name: Income Cluster, dtype: int64
```

In [57]:

```
clustering1.inertia_
```

Out[57]:

```
23517.330930930937
```

In [58]:

```
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 [59]:

```
inertia_scores
```

Out[59]:

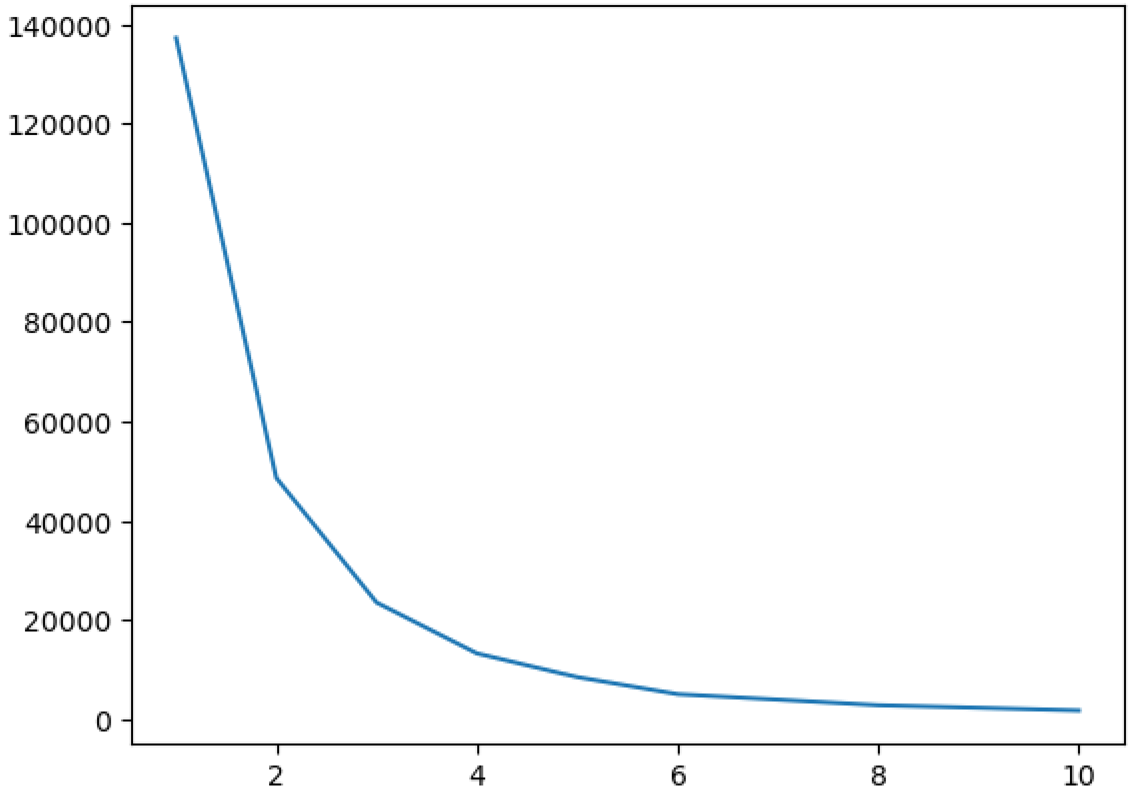
```
[137277.28,
 48660.88888888889,
 23528.152173913044,
 13278.112713472485,
 8481.496190476191,
 5050.904761904762,
 3949.2756132756135,
 2822.4996947496948,
 2304.6105580693816,
 1767.6406204906207]
```

In [60]:

```
plt.plot(range(1,11), inertia_scores)
```

Out[60]:

```
[<matplotlib.lines.Line2D at 0x253d85bda60>]
```



In [61]:
```
df.columns
```

Out[61]:
```
Index(['Gender', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)',
       'Income Cluster'],
      dtype='object')
```

In [62]:
```
df.groupby('Income Cluster')['Age', 'Annual Income (k$)', 'Spending Score (1-100)'].me
```

Out[62]:

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

Bivariate Clustering

In [68]:
```
clustering2 = KMeans(n_clusters=5)
```

In [69]:
```
clustering2.fit(df[['Annual Income (k$)', 'Spending Score (1-100)']])
clustering2.labels_
df['Spending and Income Cluster'] = clustering2.labels_
df.head()
```

Out[69]:

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

In [70]:

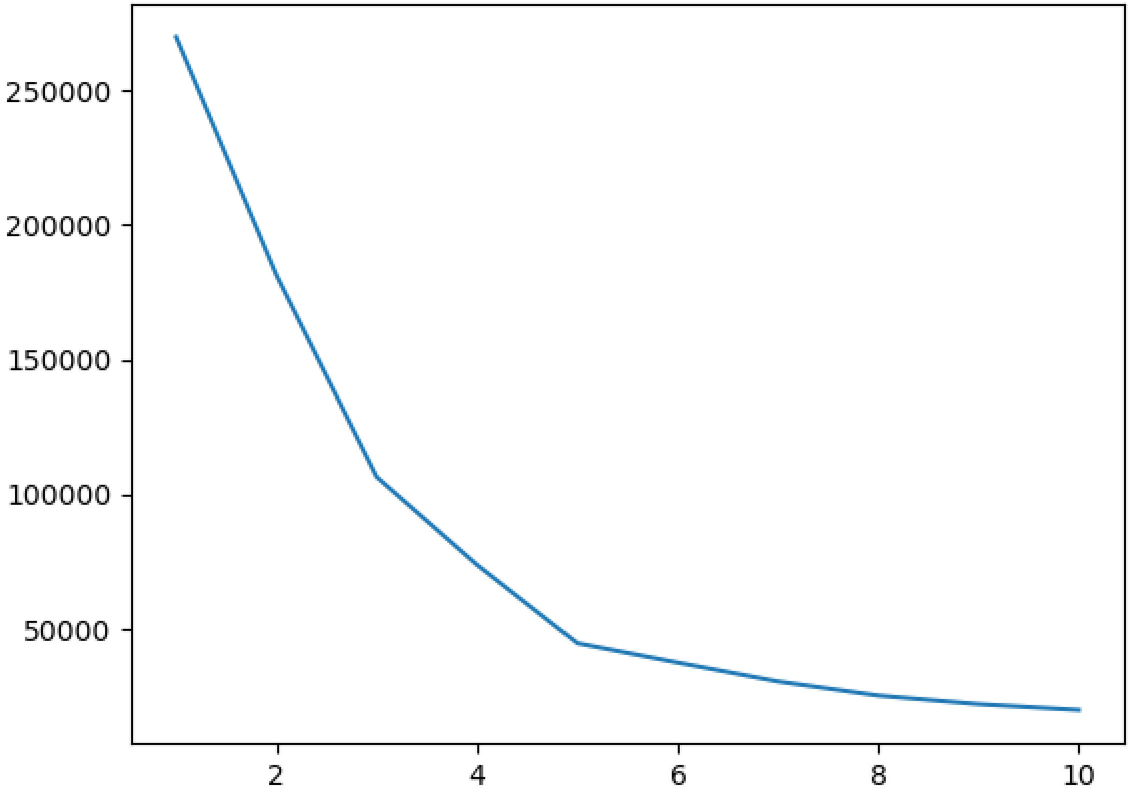
```
inertia_scores2=[]
for i in range(1,11):
    kmeans2 = KMeans(n_clusters=i)
    kmeans2.fit(df[['Annual Income (k$)', 'Spending Score (1-100)']])
    inertia_scores2.append(kmeans2.inertia_)
```

In [71]:

```
plt.plot(range(1,11), inertia_scores2)
```

Out[71]:

```
[<matplotlib.lines.Line2D at 0x253d86458e0>]
```



In [77]:

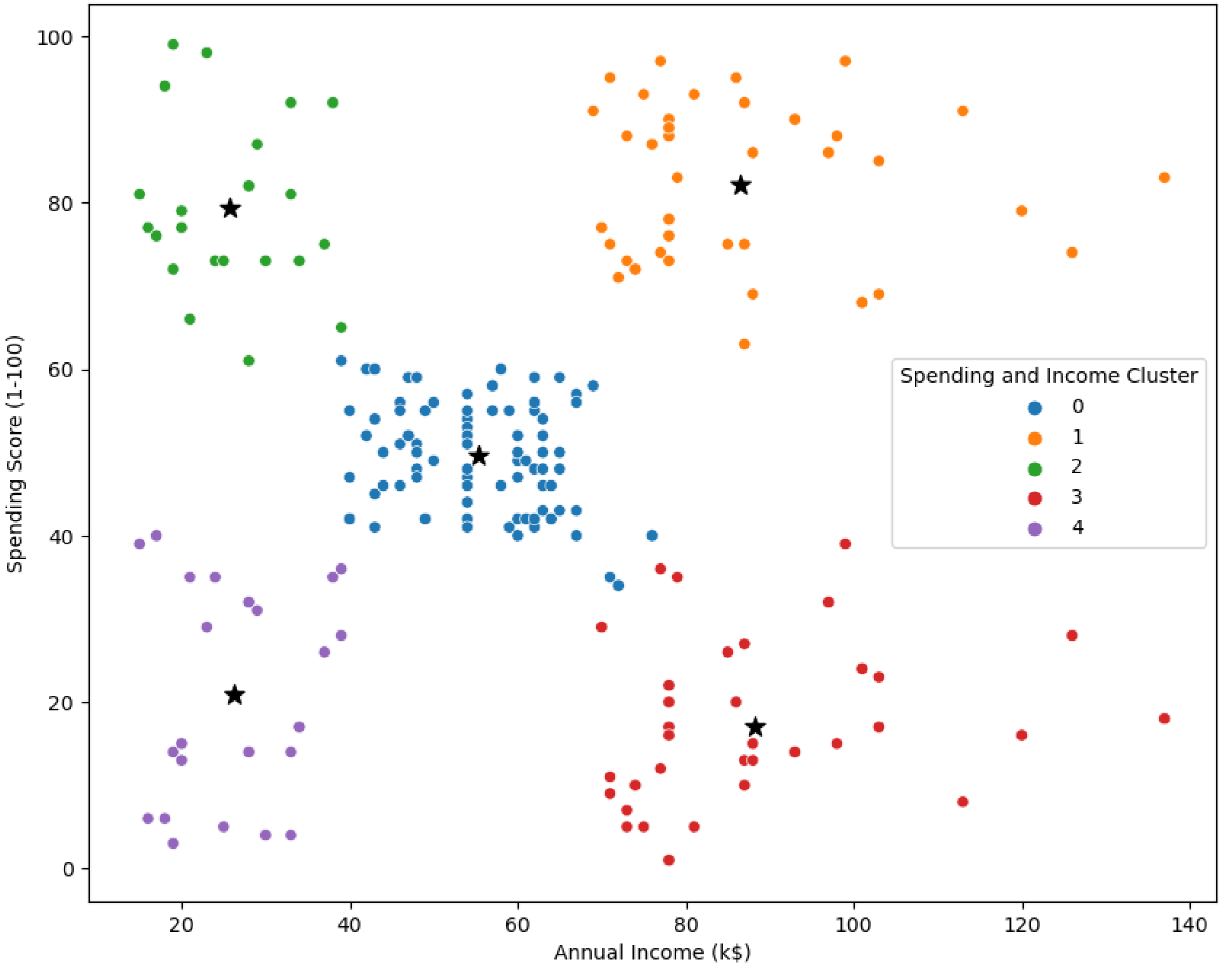
```
centers= pd.DataFrame(clustering2.cluster_centers_)
centers.columns = ['x', 'y']
```

In [78]:

```
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='Sper
```

Out[78]:

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



In [80]:
```
pd.crosstab(df['Spending and Income Cluster'], df['Gender'], normalize='index')
```

Out[80]:

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

In [81]:
```
df.groupby(['Spending and Income Cluster'])['Age', 'Annual Income (k$)',
       'Spending Score (1-100)'].mean()
```

Out[81]:

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