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('E:\Data Analyst\Python Projects\Mall\Mall_Customer.csv')
```

In [3]:

```
df.head(5)
```

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

Univariant Analysis

In [4]:

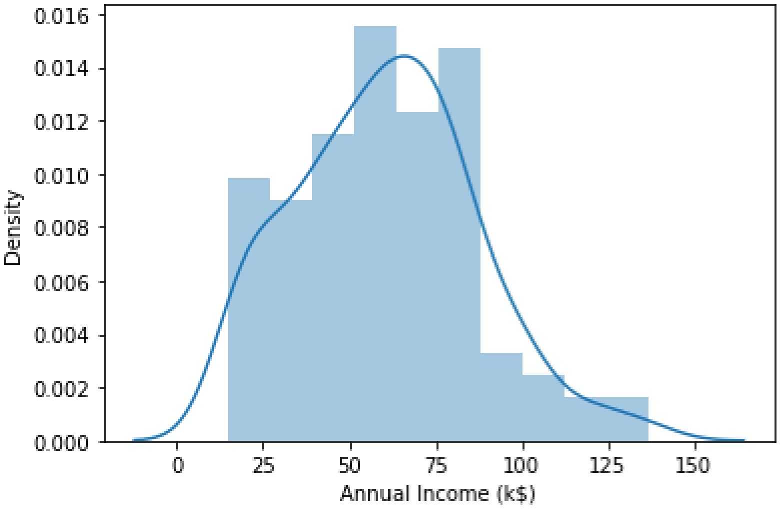
```
df.describe()
```

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]:

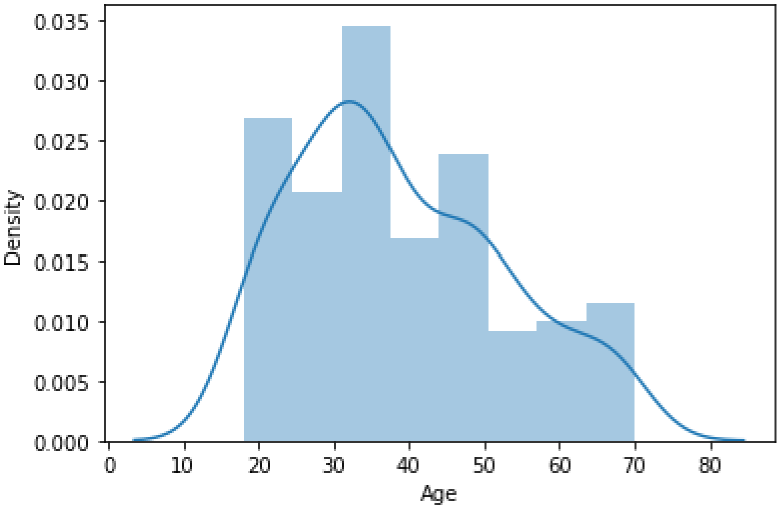
```
df.columns
```

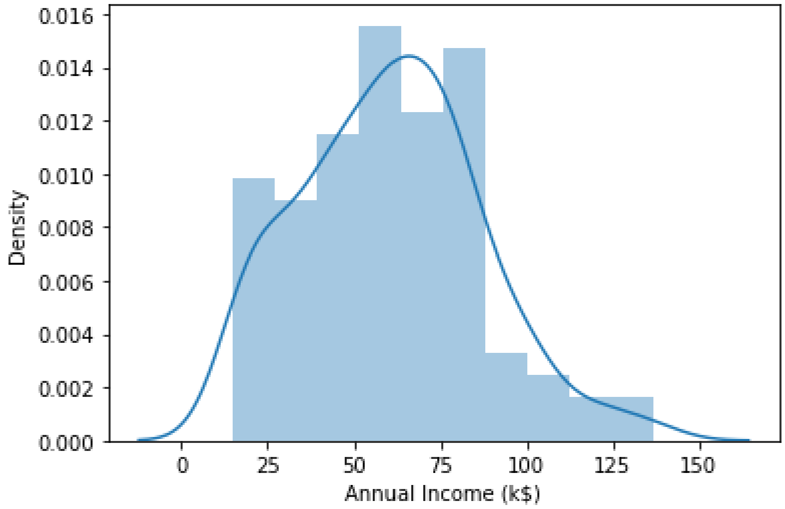
Out[6]:

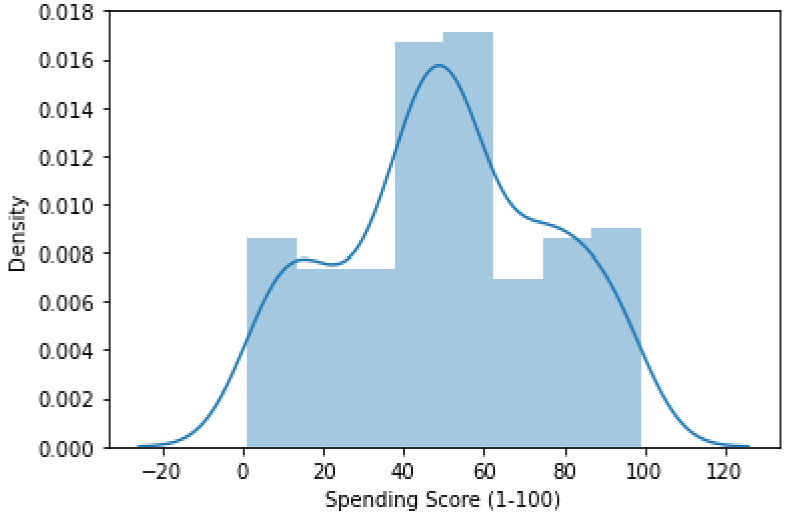
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)'],
      dtype='object')
```

In [7]:

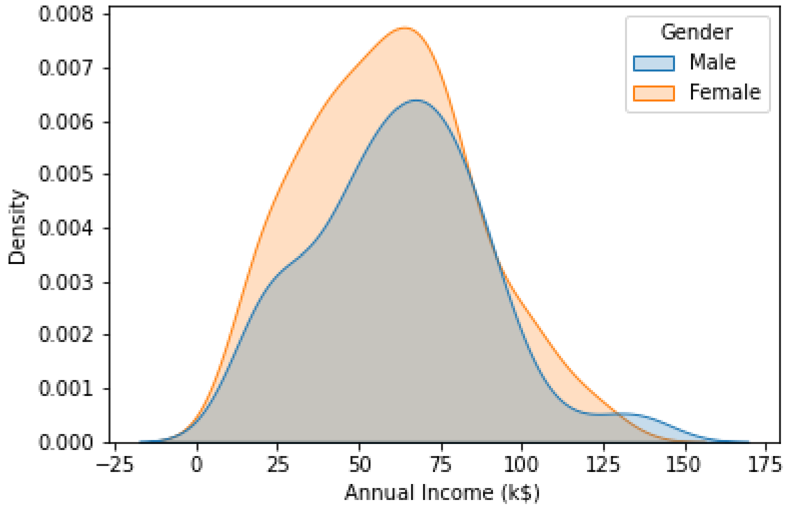
```
cols = [ 'Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in cols:
    plt.figure()
    sns.distplot(df[i])
```



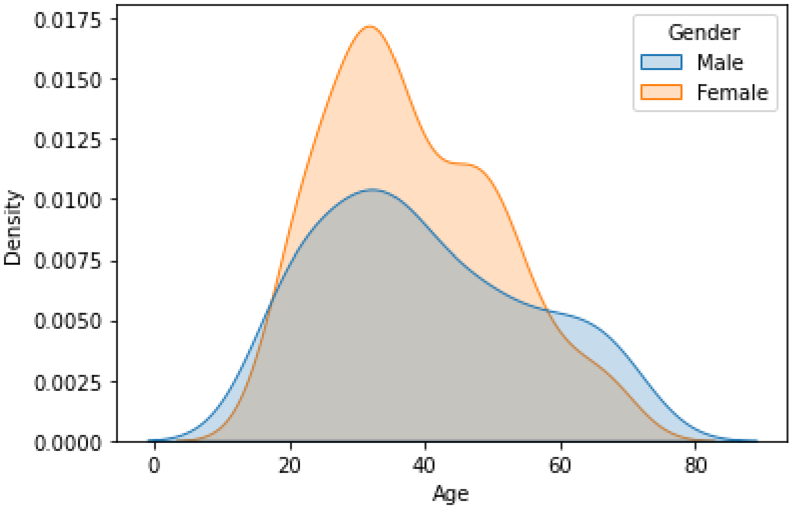


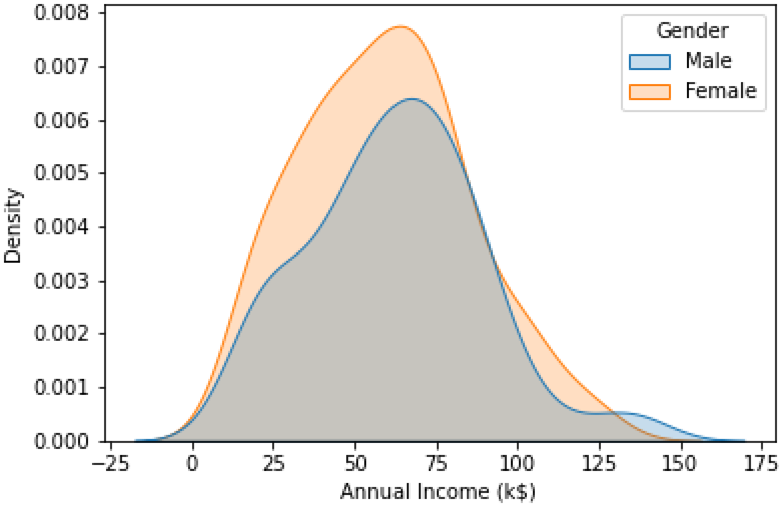


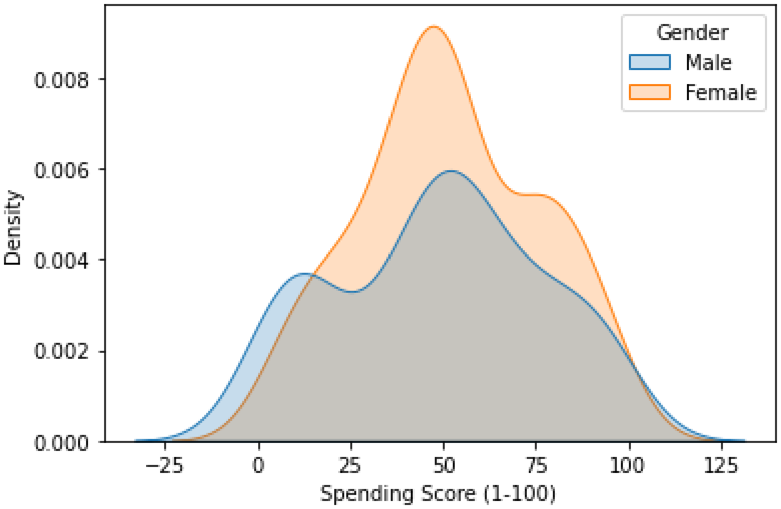
In [8]:
```
sns.kdeplot(df['Annual Income (k$)'],shade=True,hue= df['Gender']);
```



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

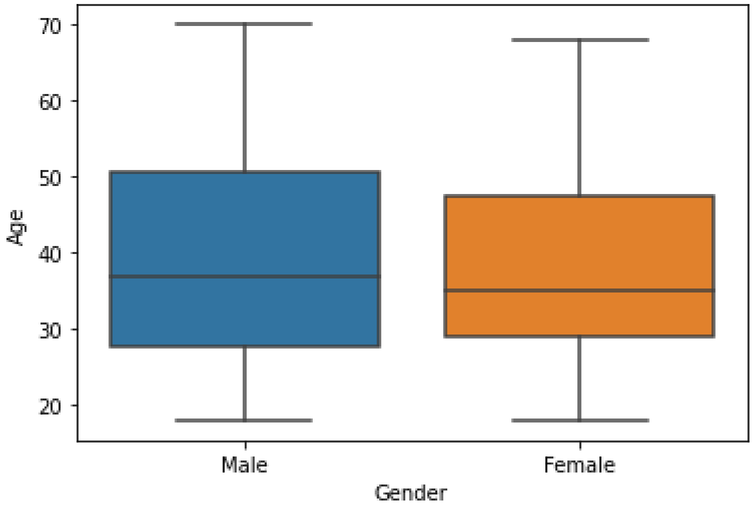


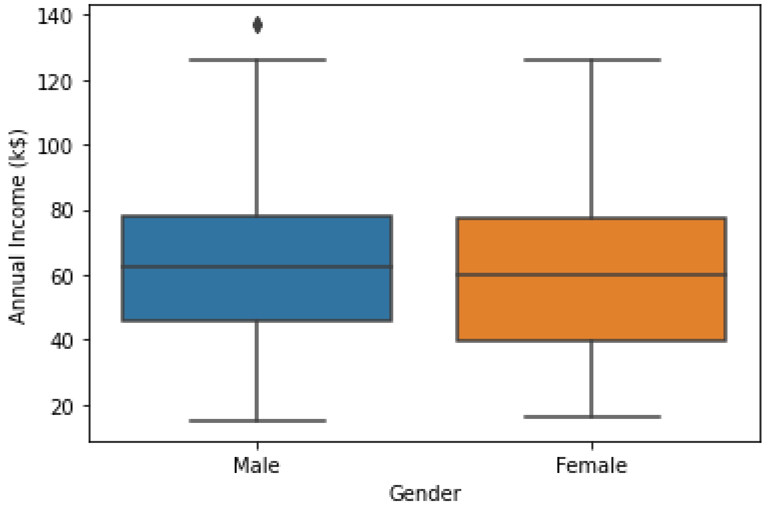


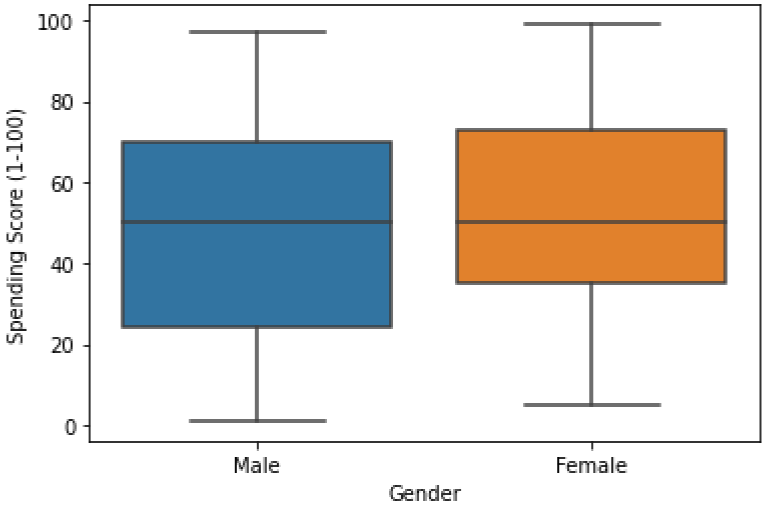


In [10]:

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







In [11]:

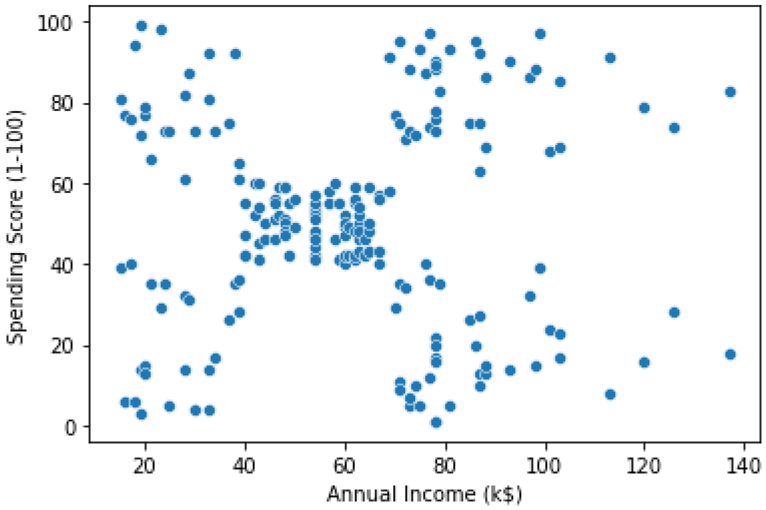
```
df['Gender'].value_counts(normalize = True)
```

Out[11]:

```
Female    0.56
Male      0.44
Name: Gender, dtype: float64
```

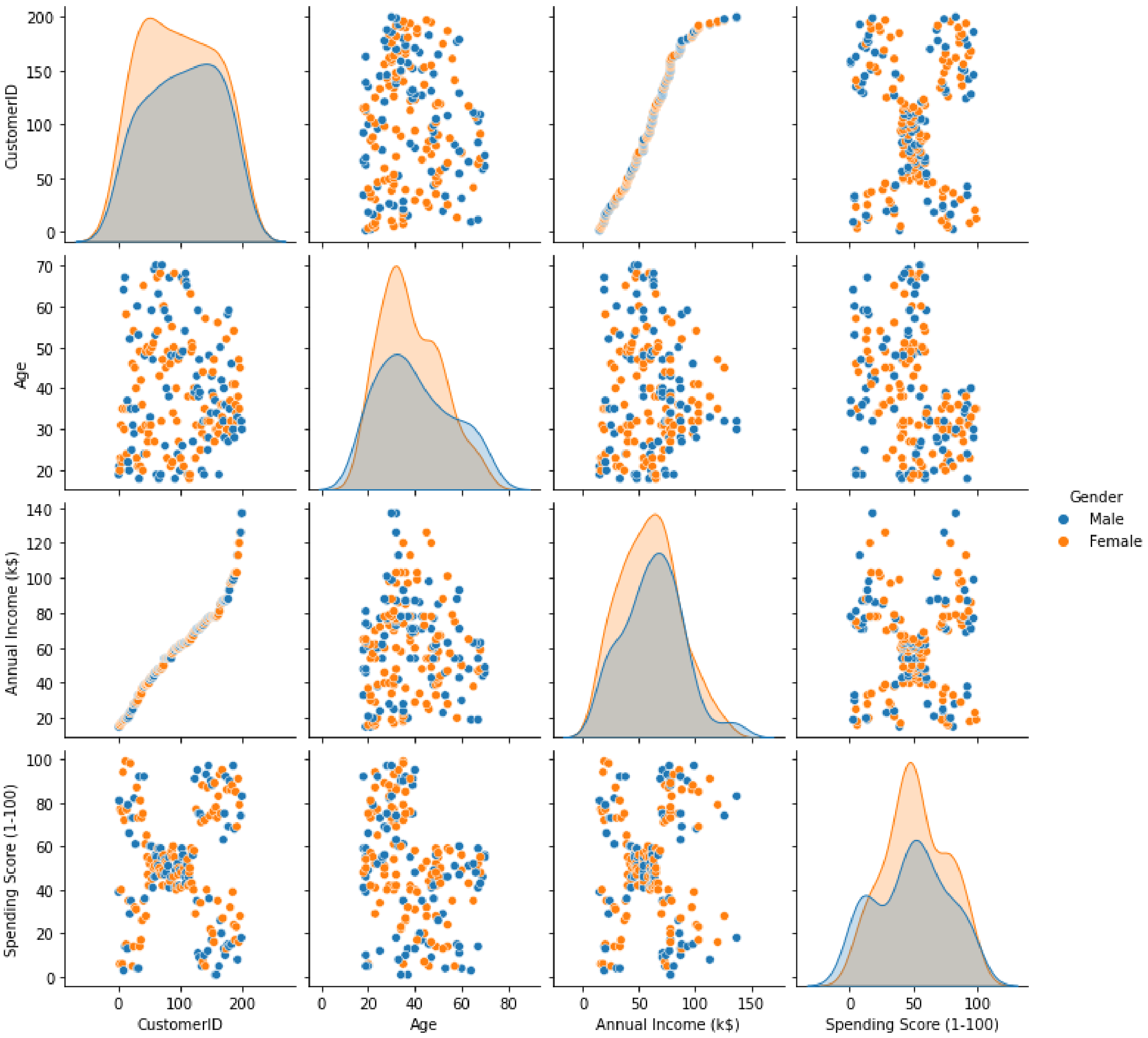
Bivariant Analysis

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



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

```
Out[13]: <seaborn.axisgrid.PairGrid at 0x1e81e003580>
```



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

Out[14]:

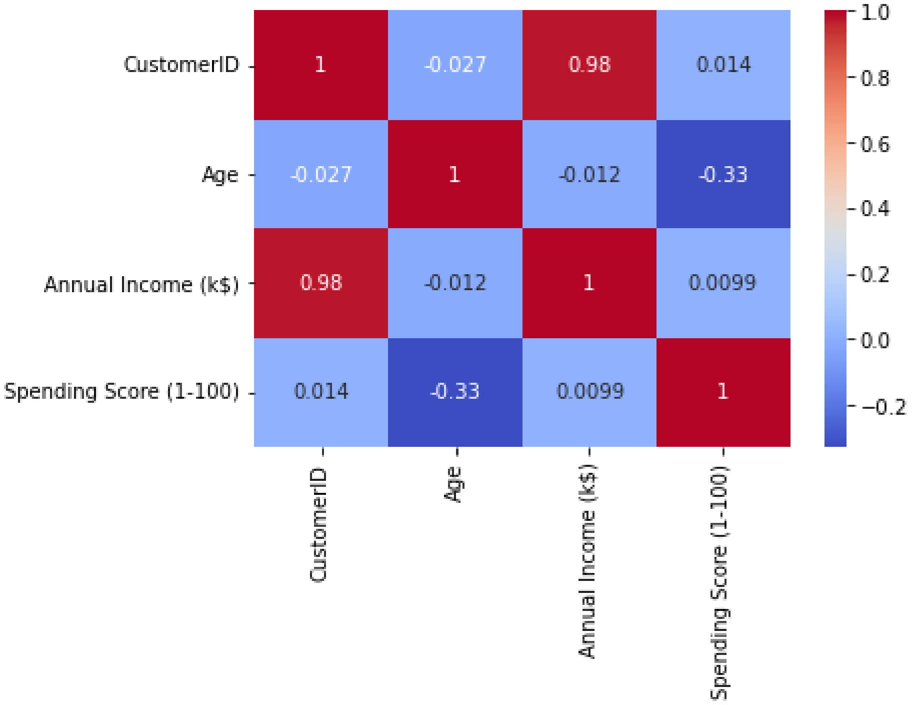
	Age	Annual Income (k$)	Spending Score (1-100)
Gender			
Female	38.098214	59.250000	51.526786
Male	39.806818	62.227273	48.511364

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');
```



Clustering - Univariant, Bivariant, Multivariant

In [17]:
```
clustering1 = KMeans(n_clusters = 3)
```

In [18]:
```
clustering1.fit(df[['Annual Income (k$)']])
```

Out[18]:
```
KMeans(n_clusters=3)
```

In [19]:
```
clustering1.labels_
```

Out[19]:

```
array([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, 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, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 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 [20]:

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

Out[20]:

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

In [21]:

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

Out[21]:

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

In [22]:

```
clustering1.inertia_
```

Out[22]:

```
23517.330930930937
```

In [23]:

```
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 [24]:

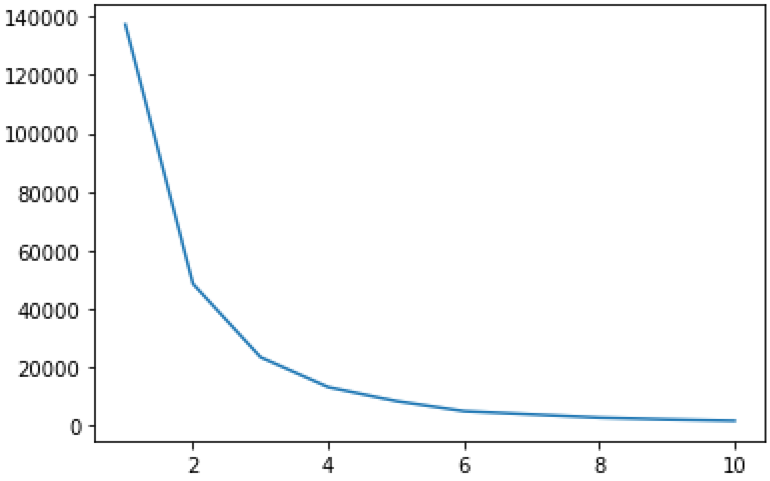
```
inertia_scores
```

Out[24]:

```
[137277.28,
 48660.88888888889,
 23517.330930930937,
 13278.112713472485,
 8493.229304029304,
 5050.904761904762,
 3931.9880952380954,
 2836.3399877899883,
 2229.5897047397048,
 1758.812049062049]
```

In [25]:

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



In [26]:
```
df.columns
```

Out[26]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)', 'Income Cluster'],
      dtype='object')
```

In [27]:
```
df.groupby('Income Cluster')['Age', 'Annual Income (k$)', 'Spending Score (1-100)'].me
```

Out[27]:

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

Bivariant Clustering

In [28]:
```
clustering2 = KMeans(n_clusters = 5)
clustering2.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
clustering2.labels_
df['Spending and Income Cluster'] = clustering2.labels_
df.head()
```

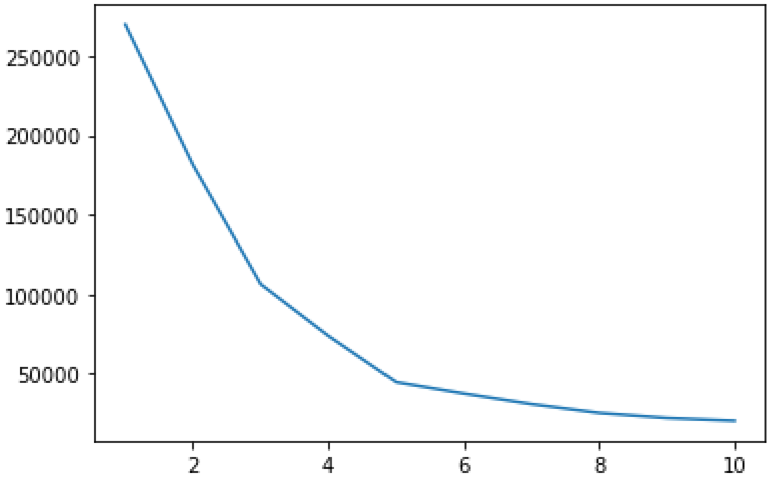
Out[28]:

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

In [29]:
```
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_)

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

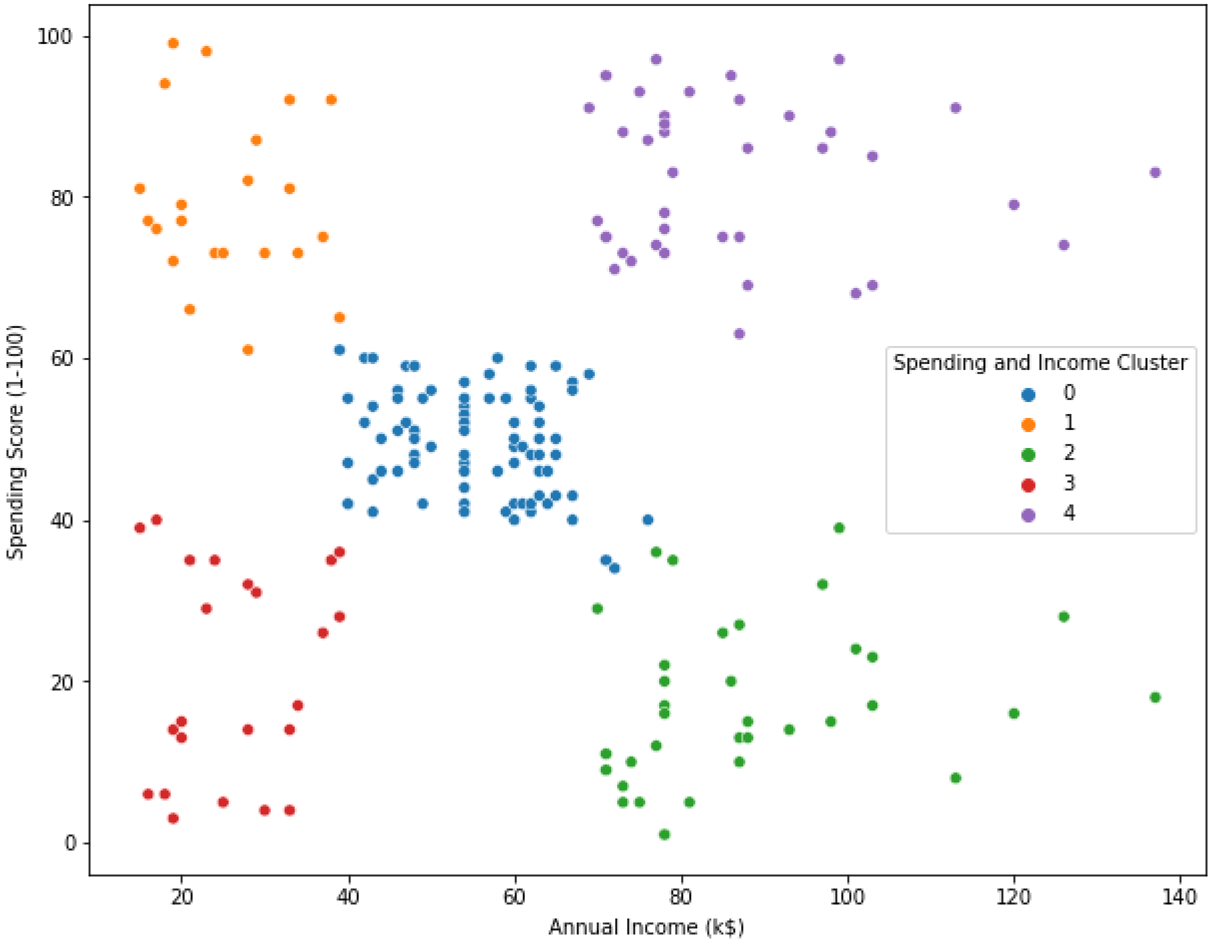


In [30]:

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

In [31]:

```
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
plt.savefig('E:\Data Analyst\Python Projects\Mall\Clustering_BIvariant.png')
```



```
In [32]: pd.crosstab(df['Spending and Income Cluster'],df['Gender'],normalize ='index')
```

Out[32]:

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

```
In [33]: df.groupby('Spending and Income Cluster')['Age', 'Annual Income (k$)', 'Spending Score
```

Out[33]:

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

Multivariant Cluster

In [34]:
```python
from sklearn.preprocessing import StandardScaler
```

In [35]:
```python
scale = StandardScaler()
```

In [36]:
```python
df.head()
```

Out[36]:

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

In [37]:
```python
dff = pd.get_dummies(df,drop_first = True)
dff.head()
```

Out[37]:

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

In [38]:
```python
dff.columns
```

Out[38]:
```
Index(['CustomerID', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)',
       'Income Cluster', 'Spending and Income Cluster', 'Gender_Male'],
      dtype='object')
```

In [39]:
```python
dff = dff[['Age', 'Annual Income (k$)', 'Spending Score (1-100)', 'Gender_Male']]
```

```
dff.head()
```

Out[39]:

	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

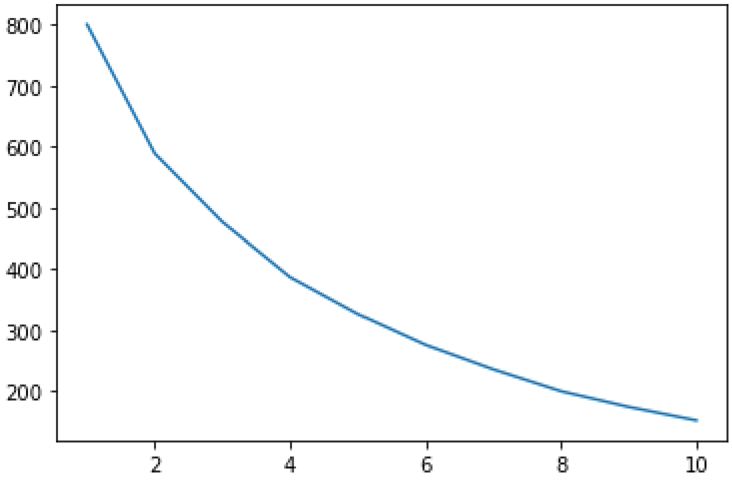
In [40]:

```
dff = pd.DataFrame(scale.fit_transform(dff))
```

In [41]:

```
inertia_scores3 = []
for i in range(1,11):
    kmeans3 = KMeans(n_clusters = i)
    kmeans3.fit(dff)
    inertia_scores3.append(kmeans3.inertia_)

plt.plot(range(1,11),inertia_scores3);
```



In [42]:

```
df.to_csv('E:\Data Analyst\Python Projects\Mall\Clustering.csv')
```

In [ ]: