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:\Customers Data.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 |

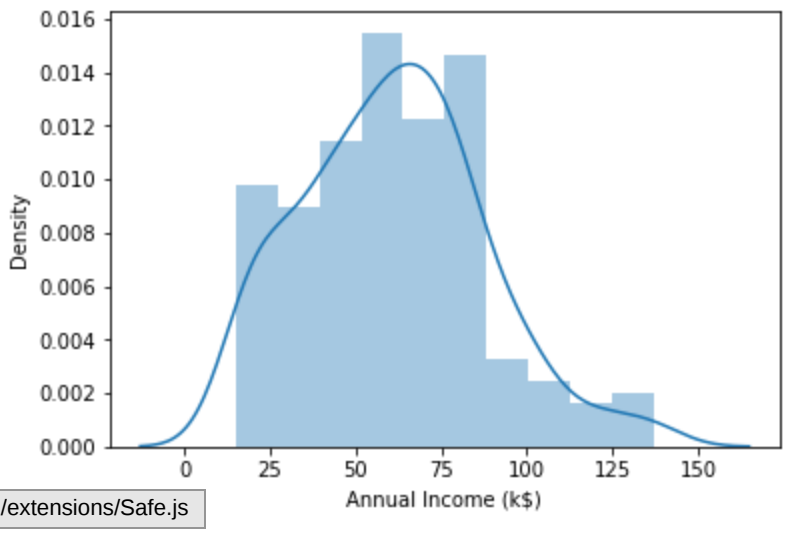
# Univariate Analysis

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

Out[4]:

| | CustomerID | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| **count** | 201.000000 | 201.000000 | 201.000000 | 201.000000 |
| **mean** | 101.000000 | 38.815920 | 60.940299 | 50.039801 |
| **std** | 58.167861 | 13.942415 | 26.748017 | 25.858817 |
| **min** | 1.000000 | 18.000000 | 15.000000 | 1.000000 |
| **25%** | 51.000000 | 29.000000 | 42.000000 | 34.000000 |
| **50%** | 101.000000 | 36.000000 | 62.000000 | 50.000000 |
| **75%** | 151.000000 | 49.000000 | 78.000000 | 73.000000 |
| **max** | 201.000000 | 70.000000 | 137.000000 | 99.000000 |

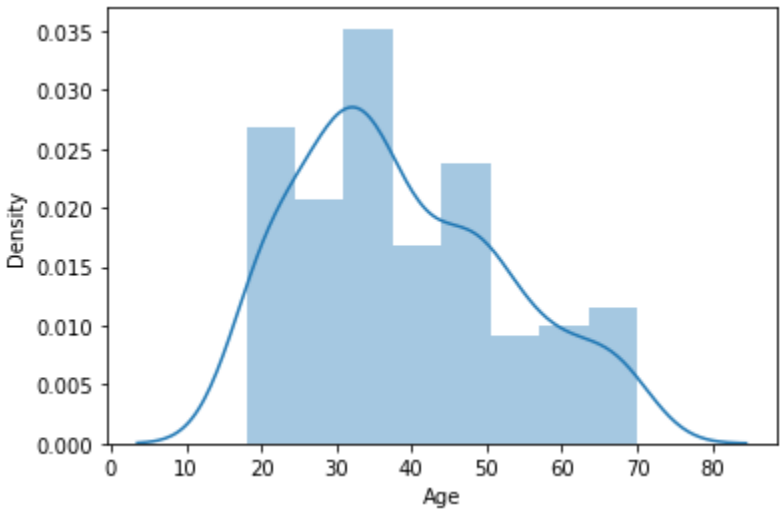
In [6]:
```
sns.distplot(df['Annual Income (k$)']);
```

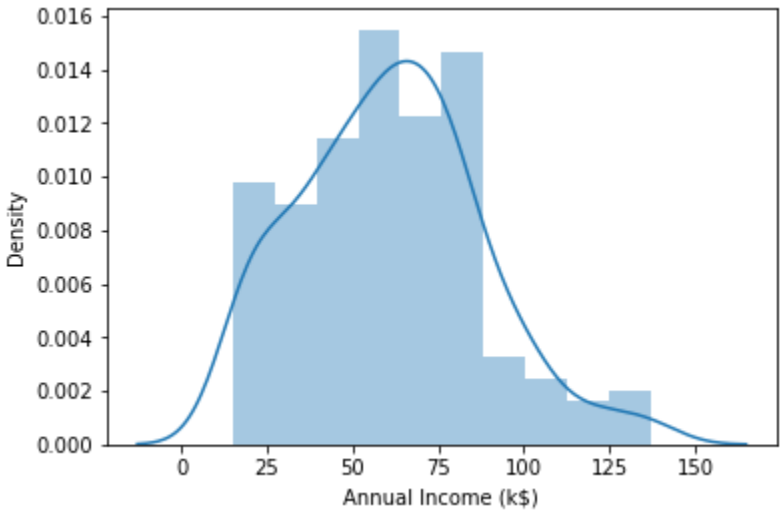


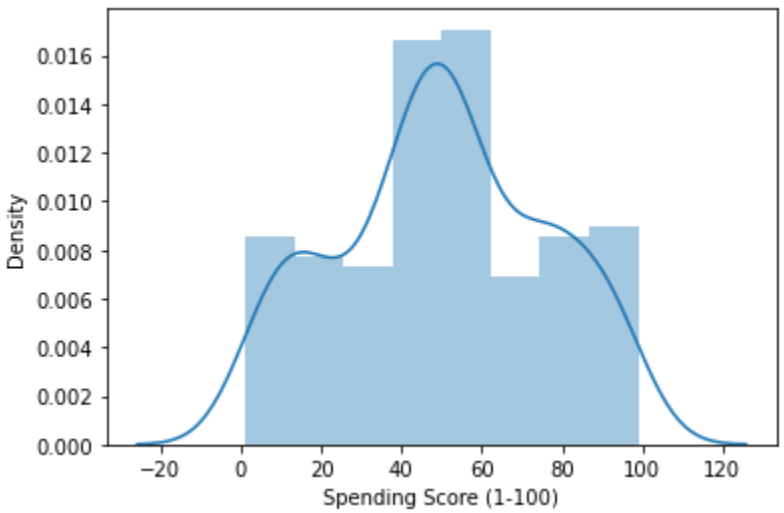
In [7]:
```
df.columns
```

Out[7]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)'],
      dtype='object')
```

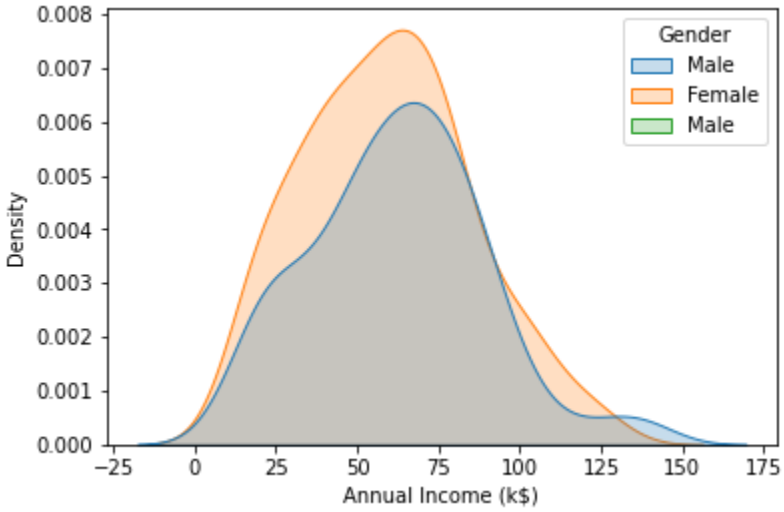
In [8]:
```
columns = ['Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.distplot(df[i])
```



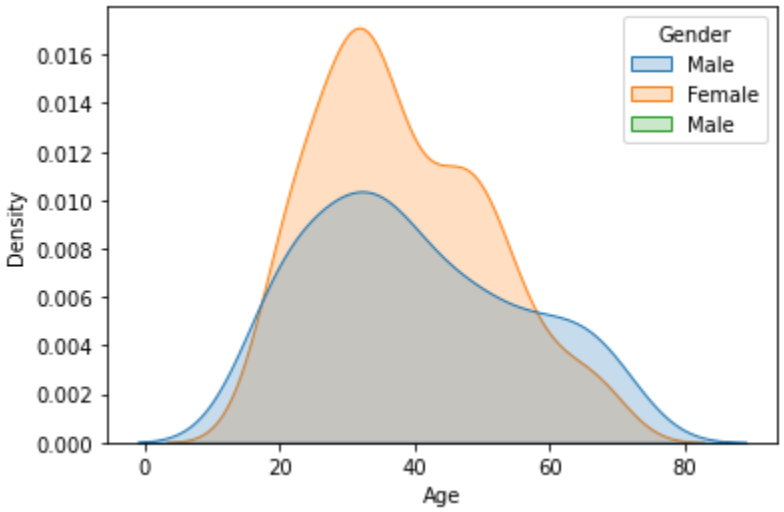



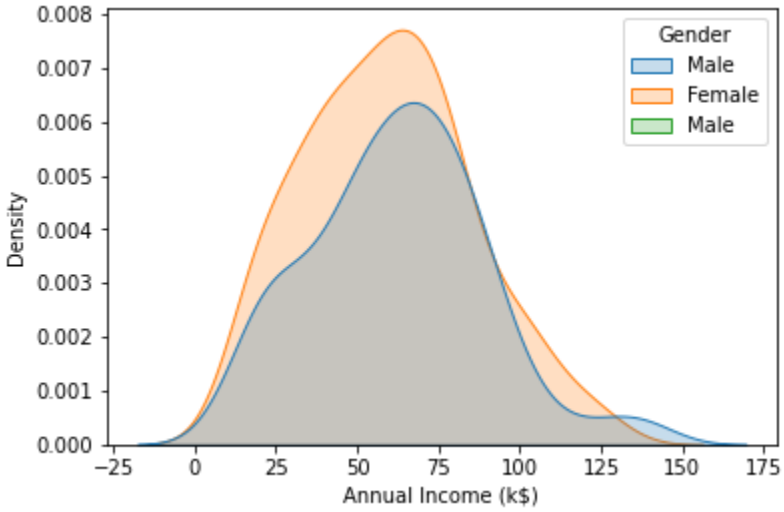



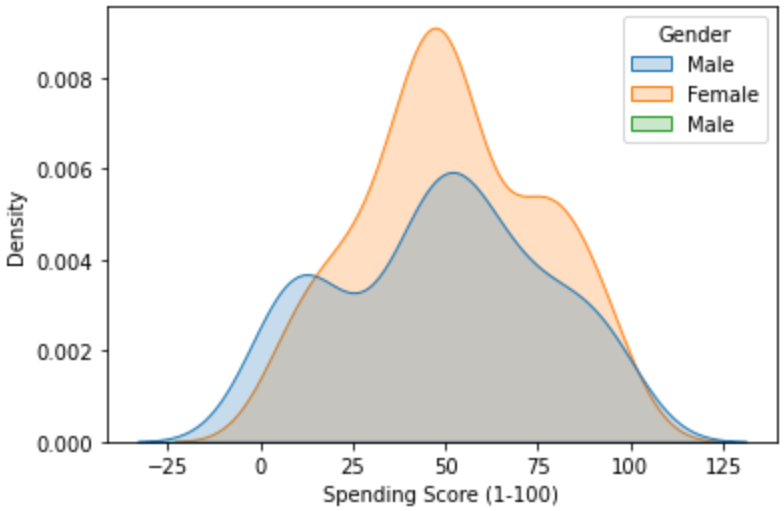

In [9]:
```
sns.kdeplot(df['Annual Income (k$)'],shade=True,hue=df['Gender']);
```



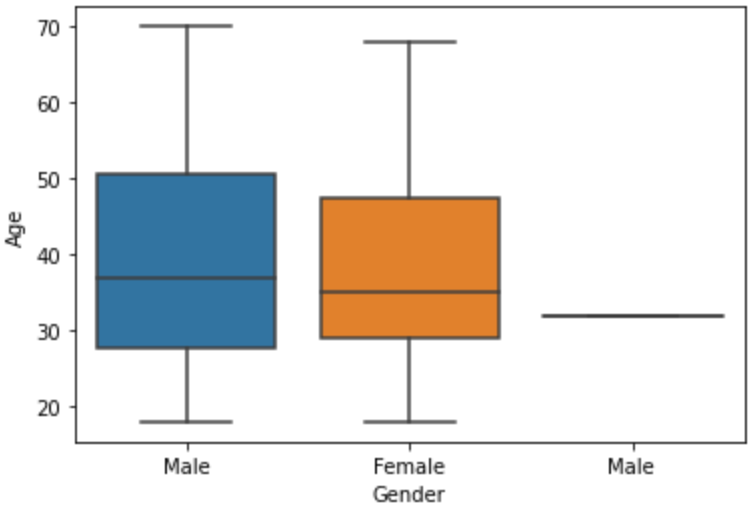

In [10]:
```
columns = ['Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.kdeplot(df[i],shade=True,hue=df['Gender'])
```

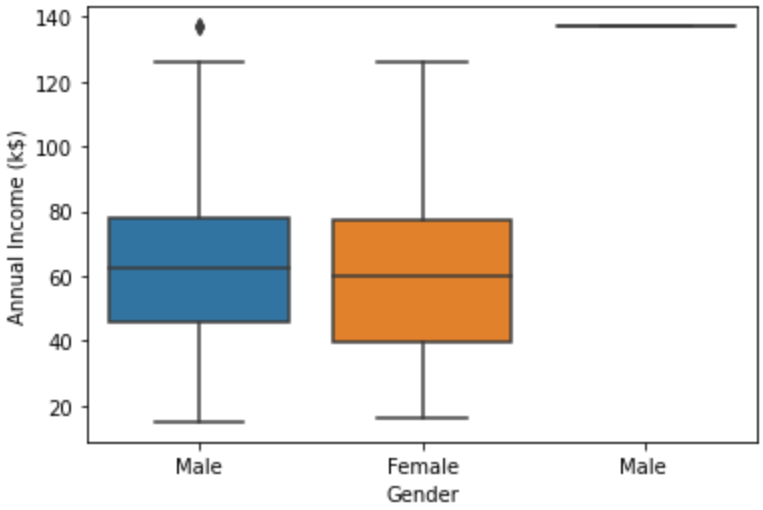


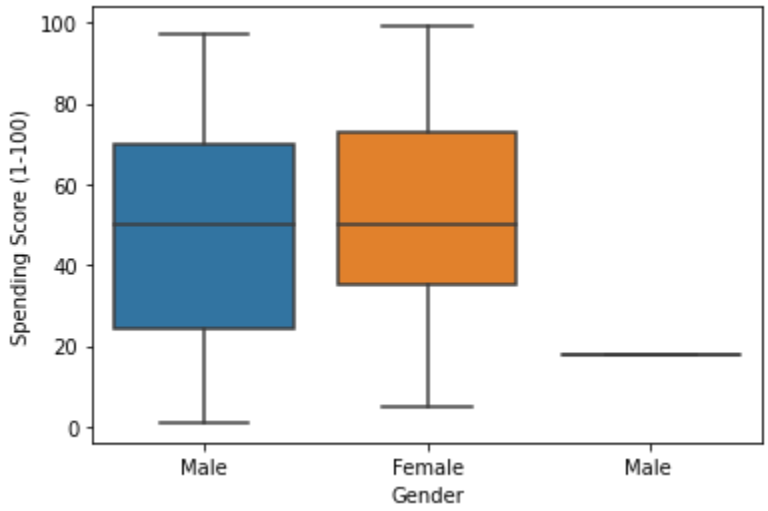






In [11]:
```
columns = ['Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.boxplot(data=df,x='Gender',y=df[i])
```





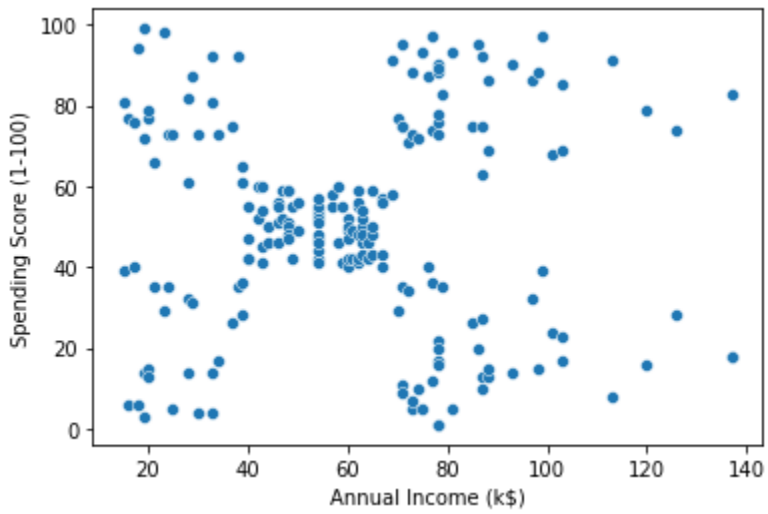

In [12]:
```
df['Gender'].value_counts(normalize=True)
```

Out[12]:
```
Female    0.557214
Male      0.437811
Male      0.004975
Name: Gender, dtype: float64
```

# Bivariate Analysis

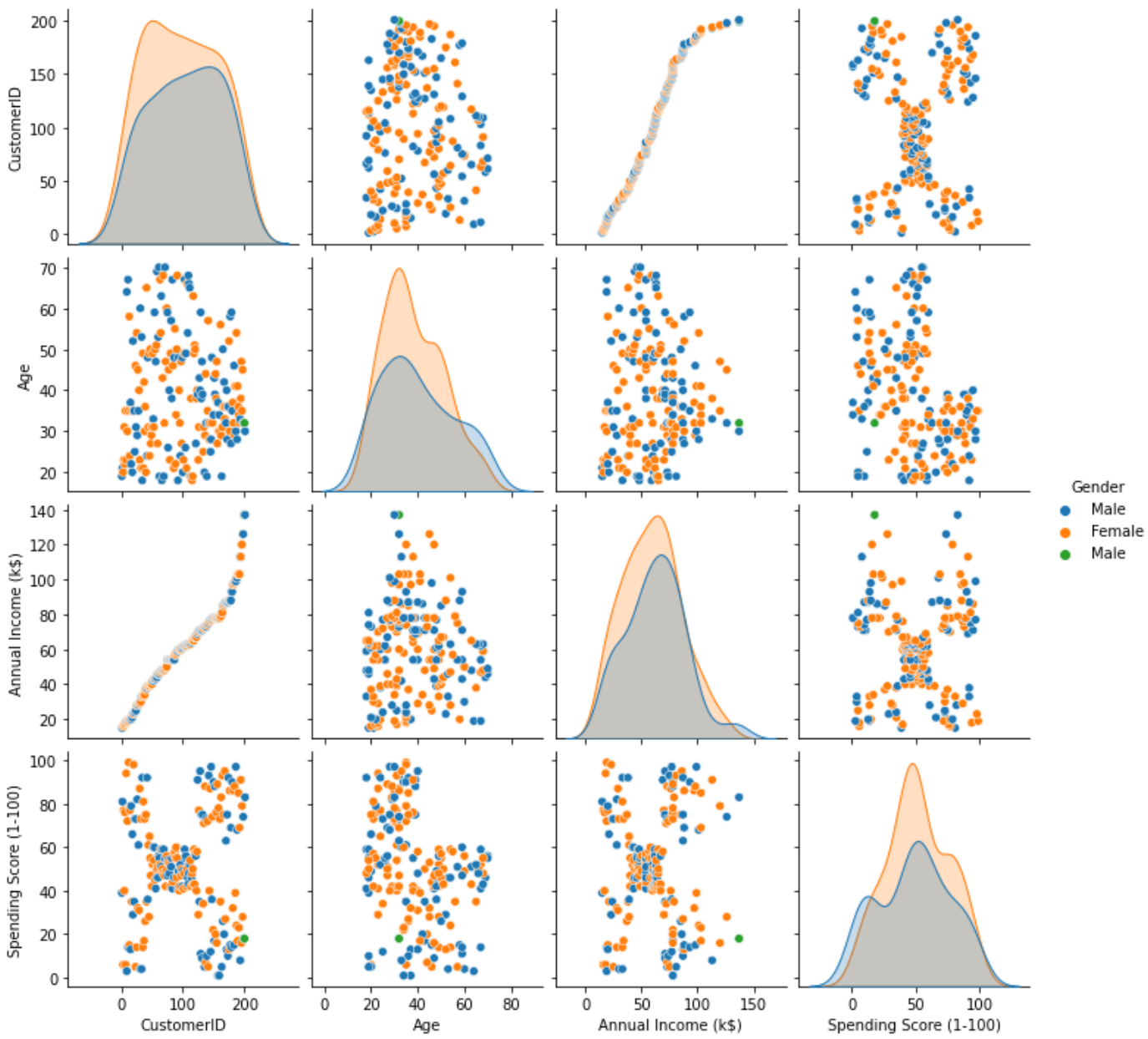
In [13]:
```
sns.scatterplot(data=df, x='Annual Income (k$)',y='Spending Score (1-100)' )
```

Out[13]:
```
<AxesSubplot:xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'>
```




In [14]:
```
#df=df.drop('CustomerID',axis=1)
sns.pairplot(df,hue='Gender')
```

Out[14]:
```
<seaborn.axisgrid.PairGrid at 0x4cb1033f40>
```




In [15]:
```
df.groupby(['Gender'])['Age', 'Annual Income (k$)',
       'Spending Score (1-100)'].mean()
```

Out[15]:

| Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|
| **Female** | 38.098214 | 59.250000 | 51.526786 |
| **Male** | 39.806818 | 62.227273 | 48.511364 |
| **Male** | 32.000000 | 137.000000 | 18.000000 |

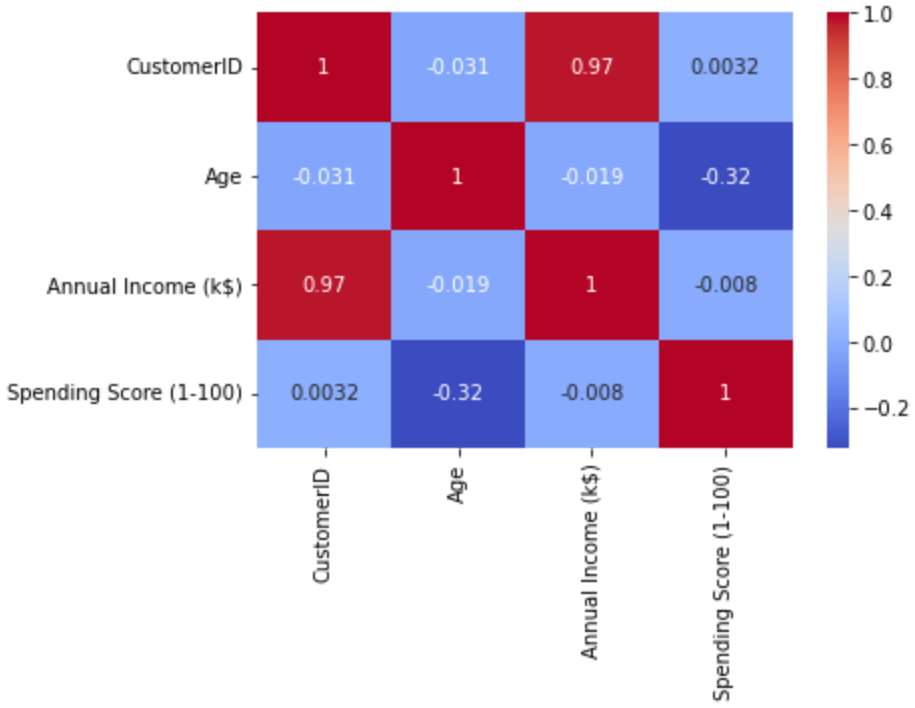
In [16]:
```
df.corr()
```

Out[16]:

| | CustomerID | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| **CustomerID** | 1.000000 | -0.030783 | 0.974911 | 0.003191 |
| **Age** | -0.030783 | 1.000000 | -0.019122 | -0.322723 |
| **Annual Income (k$)** | 0.974911 | -0.019122 | 1.000000 | -0.008042 |
| **Spending Score (1-100)** | 0.003191 | -0.322723 | -0.008042 | 1.000000 |

```
df.corr(),annot=True,cmap='coolwarm')
```

Out[17]:
```
<AxesSubplot:>
```




# Clustering - Univariate, Bivariate, Multivariate

In [18]:
```
clustering1 = KMeans(n_clusters=3)
```

In [19]:
```
clustering1.fit(df[['Annual Income (k$)']])
```

Out[19]:
```
KMeans(n_clusters=3)
```

In [20]:
```
clustering1.labels_
```

Out[20]:
```
array([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, 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, 2, 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, 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])
```

In [21]:
```
df['Income Cluster'] = clustering1.labels_
df.head()
```

Out[21]:

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

```
luster'].value_counts()
```

Out[22]:
```
0    90
2    74
1    37
Name: Income Cluster, dtype: int64
```

In [23]:
```
clustering1.inertia_
```

Out[23]:
```
24857.342942942945
```

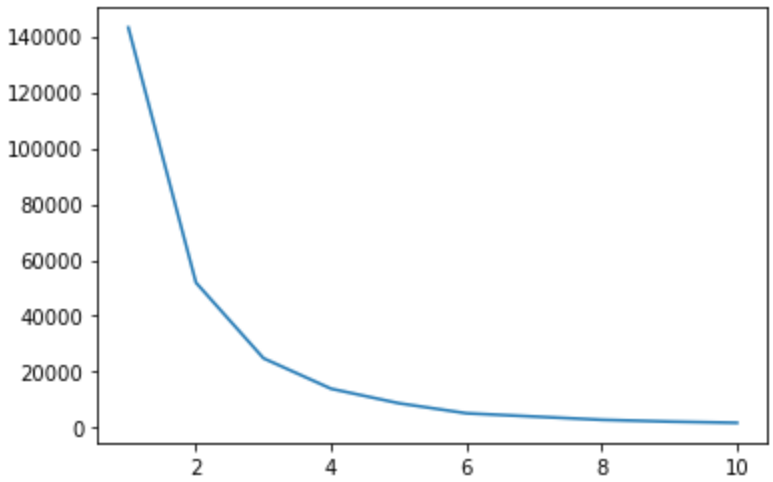
In [24]:
```
intertia_scores=[]
for i in range(1,11):
    kmeans=KMeans(n_clusters=i)
    kmeans.fit(df[['Annual Income (k$)']])
    intertia_scores.append(kmeans.inertia_)
```

In [25]:
```
intertia_scores
```

Out[25]:
```
[143091.28358208953,
 51925.96636636638,
 24868.16418592506,
 13987.912713472491,
 8817.90183705973,
 5231.70688248949,
 4059.573725981618,
 2897.1663614163595,
 2259.9551575875103,
 1821.1374458874457]
```

In [26]:
```
plt.plot(range(1,11),intertia_scores)
```

Out[26]:
```
[<matplotlib.lines.Line2D at 0x4cb74f6160>]
```




In [27]:
```
df.columns
```

Out[27]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)', 'Income Cluster'],
      dtype='object')
```

In [28]:
```
df.groupby('Income Cluster')['Age', 'Annual Income (k$)',
       'Spending Score (1-100)'].mean()
```

Out[28]:

| Income Cluster | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|
| 0 | 38.722222 | 67.088889 | 50.000000 |
| 1 | 37.675676 | 100.891892 | 49.756757 |
| 2 | 39.500000 | 33.486486 | 50.229730 |

# Bivariate Clustering

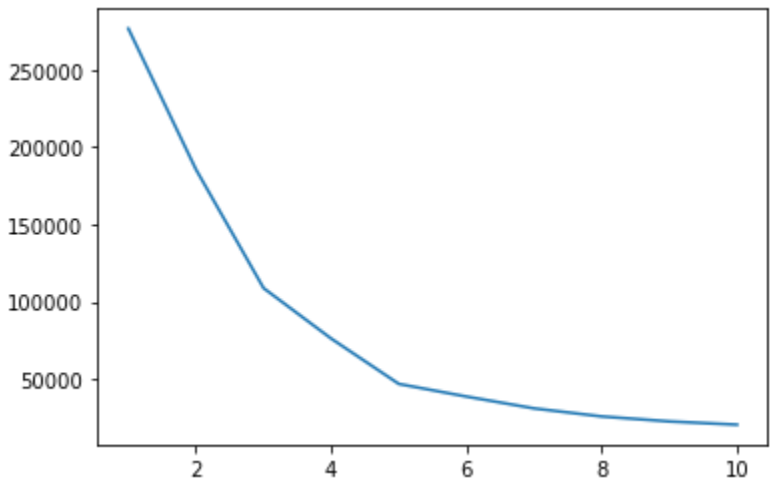
In [29]:
```
clustering2 = KMeans(n_clusters=5)
clustering2.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
df['Spending and Income Cluster'] =clustering2.labels_
df.head()
```

Out[29]:

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

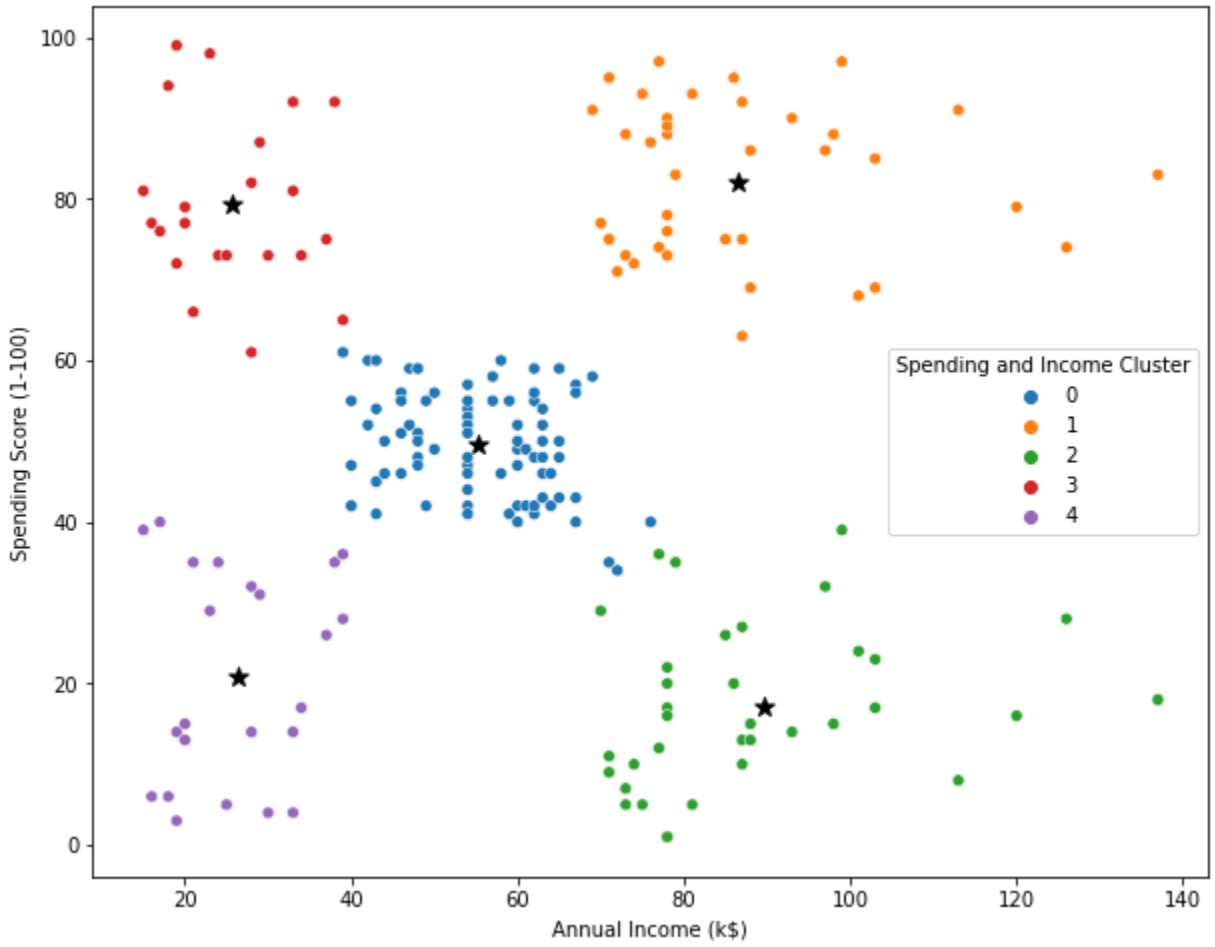
In [30]:
```
intertia_scores2=[]
for i in range(1,11):
    kmeans2=KMeans(n_clusters=i)
    kmeans2.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
    intertia_scores2.append(kmeans2.inertia_)
plt.plot(range(1,11),intertia_scores2)
```

Out[30]:
```
[<matplotlib.lines.Line2D at 0x4cb8f2cd00>]
```




In [31]:
```
centers =pd.DataFrame(clustering2.cluster_centers_)
centers.columns = ['x','y']
```

In [32]:
```
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='Spendin
plt.savefig('clustering_bivaraiate.png')
```




In [33]:
```
pd.crosstab(df['Spending and Income Cluster'],df['Gender'],normalize='index')
```

Out[33]:

| Gender | Female | Male | Male |
|---|---|---|---|
| **Spending and Income Cluster** | | | |
| **0** | 0.592593 | 0.407407 | 0.000000 |
| **1** | 0.538462 | 0.461538 | 0.000000 |
| **2** | 0.444444 | 0.527778 | 0.027778 |
| **3** | 0.590909 | 0.409091 | 0.000000 |
| **4** | 0.608696 | 0.391304 | 0.000000 |

In [34]:
```
df.groupby('Spending and Income Cluster')['Age', 'Annual Income (k$)',
       'Spending Score (1-100)'].mean()
```

Out[34]:

| | 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** | 40.861111 | 89.555556 | 17.138889 |
| **3** | 25.272727 | 25.727273 | 79.363636 |
| **4** | 45.217391 | 26.304348 | 20.913043 |

# mulivariate clustering

In [35]:
```
from sklearn.preprocessing import StandardScaler
```

In [36]:
```
scale = StandardScaler()
```

In [37]:
```
df.head()
```

Out[37]:

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

In [38]:
```
dff = pd.get_dummies(df,drop_first=True)
dff.head()
```

Out[38]:

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

In [39]:
```
dff.columns
```

Out[39]:
```
Index(['CustomerID', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)',
       'Income Cluster', 'Spending and Income Cluster', 'Gender_Male',
       'Gender_Male '],
      dtype='object')
```

In [40]:
```
dff = dff[['Age', 'Annual Income (k$)', 'Spending Score (1-100)','Gender_Male']]
dff.head()
```

Out[40]:

| | 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 [41]:
```
dff = scale.fit_transform(dff)
```

In [42]:
```
dff = pd.DataFrame(scale.fit_transform(dff))
dff.head()
```

Out[42]:

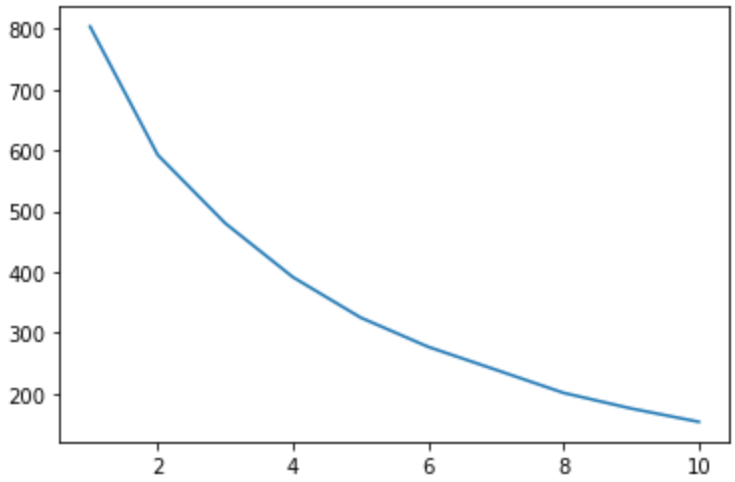
| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **0** | -1.424818 | -1.721810 | -0.427992 | 1.133177 |
| **1** | -1.281012 | -1.721810 | 1.200268 | 1.133177 |
| **2** | -1.352915 | -1.684331 | -1.707339 | -0.882474 |
| **3** | -1.137207 | -1.684331 | 1.045195 | -0.882474 |
| **4** | -0.561986 | -1.646852 | -0.389224 | -0.882474 |

In [43]:

```
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[43]:

```
[<matplotlib.lines.Line2D at 0x4cba2fdc10>]
```




In [44]:

```
df
```

Out[44]:

| | CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) | Income Cluster | Spending and Income Cluster |
|---|---|---|---|---|---|---|---|
| **0** | 1 | Male | 19 | 15 | 39 | 2 | 4 |
| **1** | 2 | Male | 21 | 15 | 81 | 2 | 3 |
| **2** | 3 | Female | 20 | 16 | 6 | 2 | 4 |
| **3** | 4 | Female | 23 | 16 | 77 | 2 | 3 |
| **4** | 5 | Female | 31 | 17 | 40 | 2 | 4 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **196** | 197 | Female | 45 | 126 | 28 | 1 | 2 |
| **197** | 198 | Male | 32 | 126 | 74 | 1 | 1 |
| **198** | 199 | Male | 32 | 137 | 18 | 1 | 2 |
| **199** | 200 | Male | 32 | 137 | 18 | 1 | 2 |
| **200** | 201 | Male | 30 | 137 | 83 | 1 | 1 |

201 rows × 7 columns

In [45]:

```
df.to_csv('Clustering.csv')
```

In [ ]: