MallMaven (1)

June 9, 2024

```
[1]: import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     import warnings
     warnings.filterwarnings('ignore')
     %pip install -Uq notebook-as-pdf
```

```
[2]: df = pd.read_csv('Mall_Customers.csv')
```

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

```
[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
```

1 Univariate Analysis

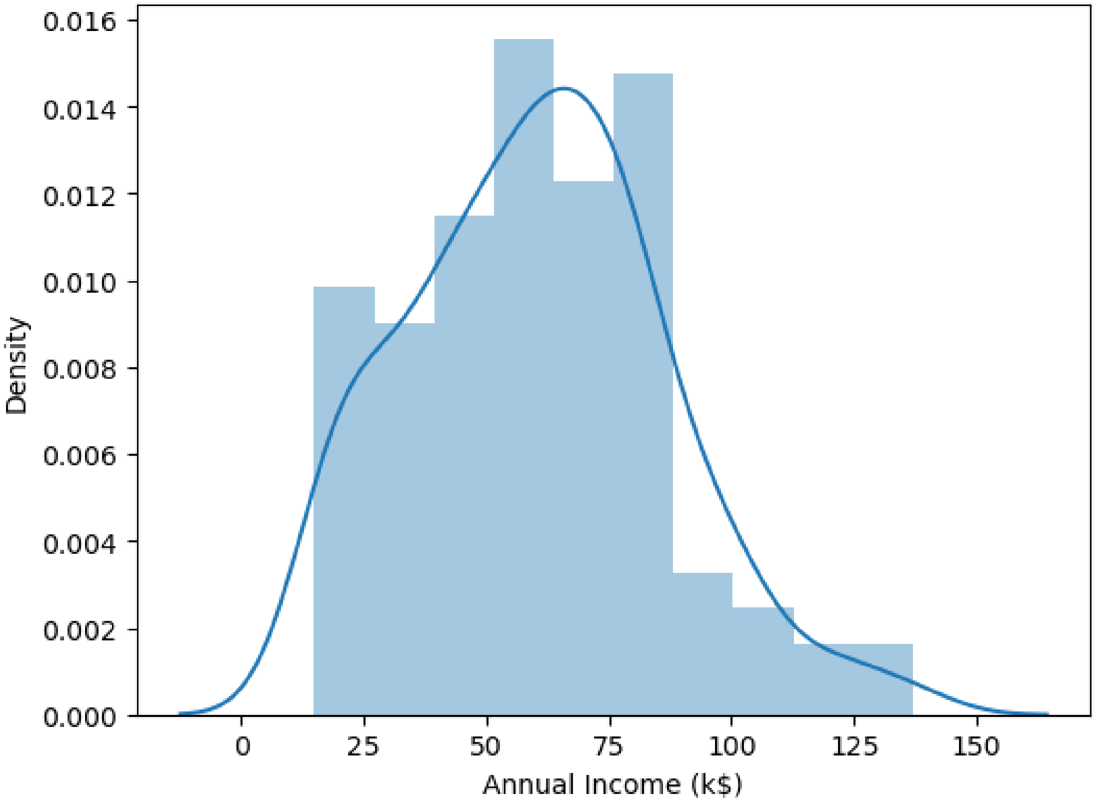
```
[4]: df = df.drop(columns=['CustomerID'])
```

```
[5]: df.describe()
```

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

Histogram to look at one of the column

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



[7]:

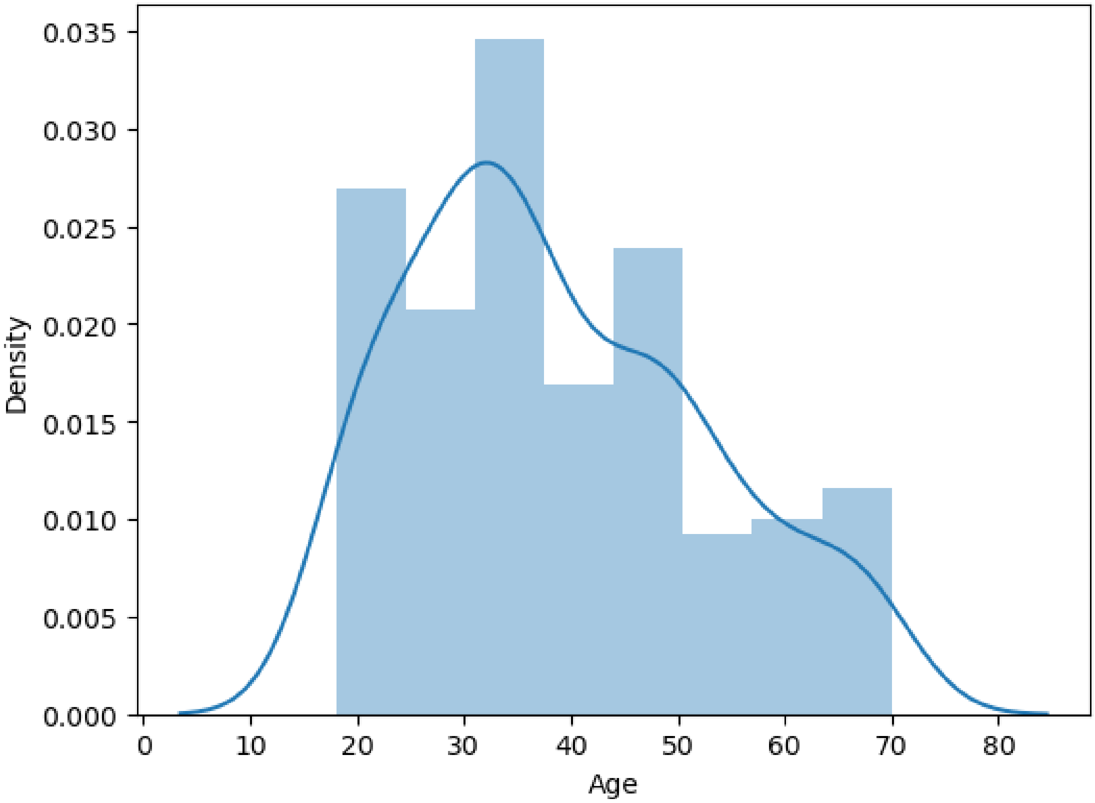
```
df.columns
```

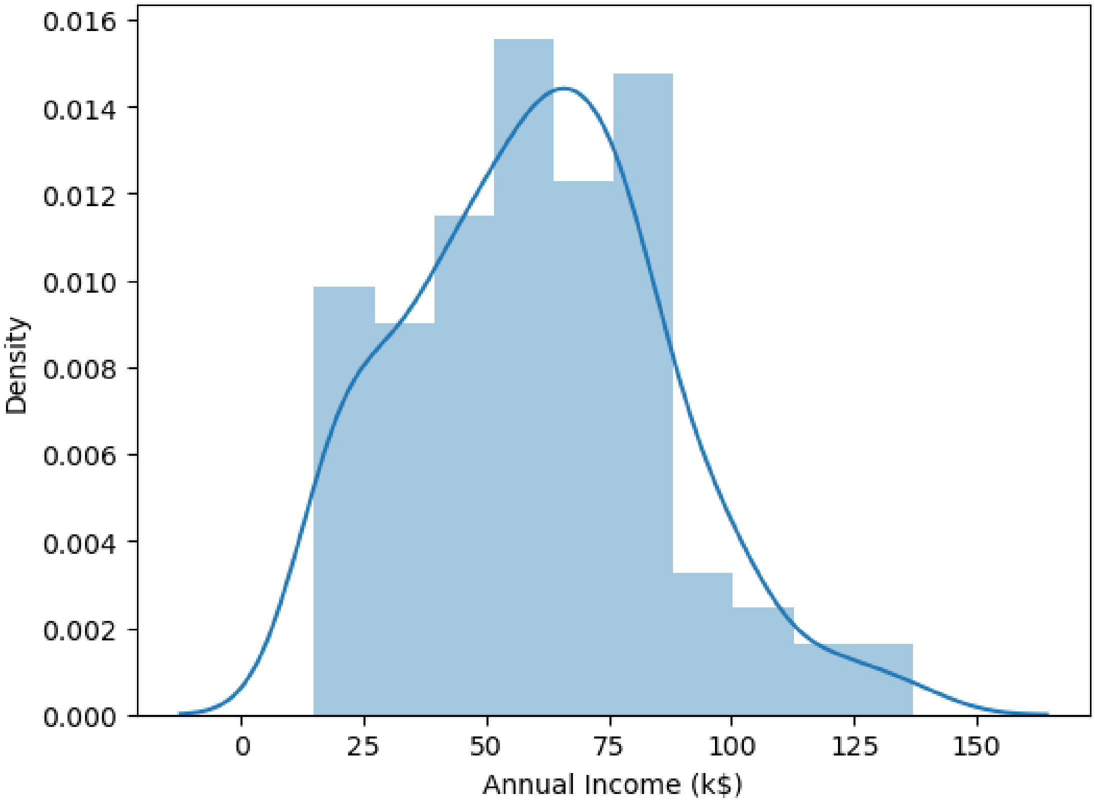
[7]:

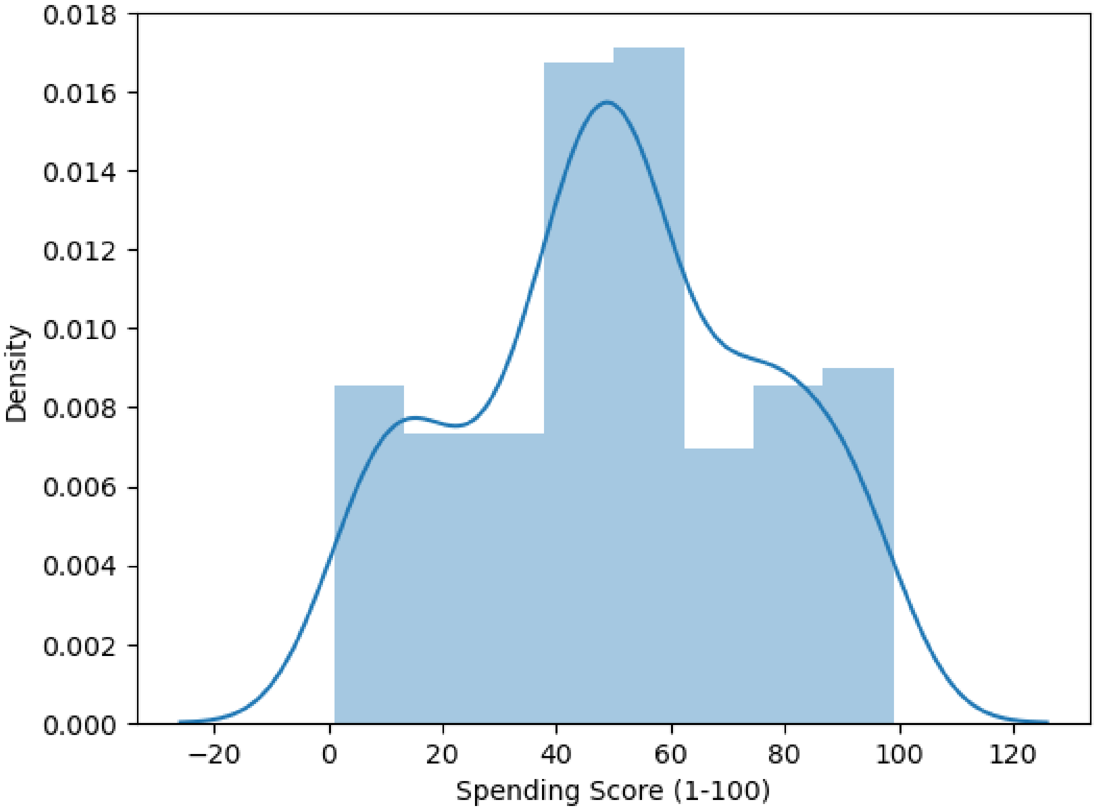
```
Index(['Gender', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)'],
dtype='object')
```

[8]:

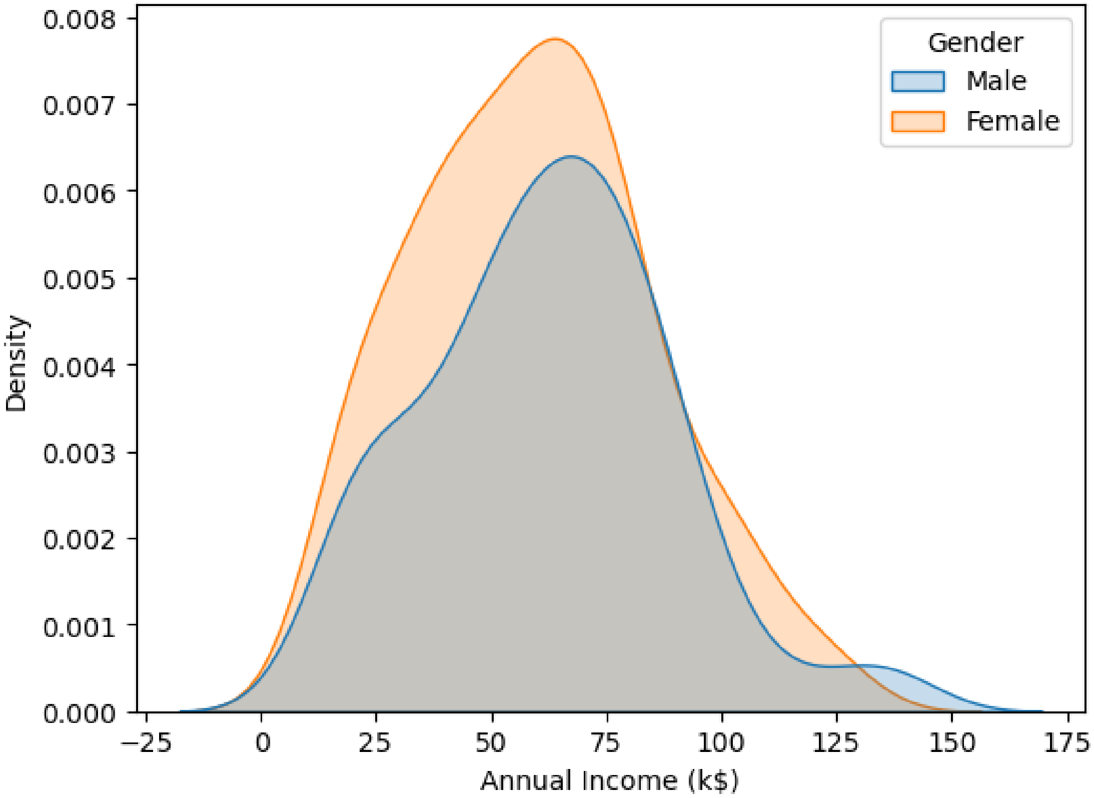
```
columns = ['Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.distplot(df[i])
```


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$)

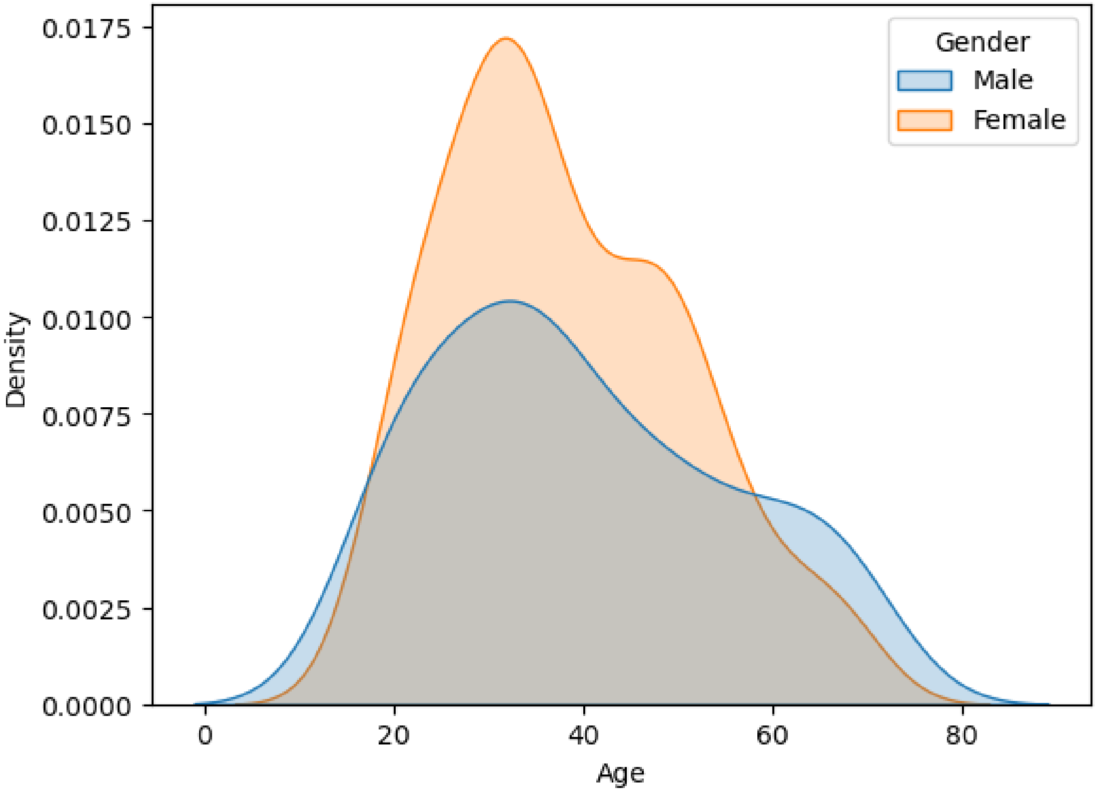


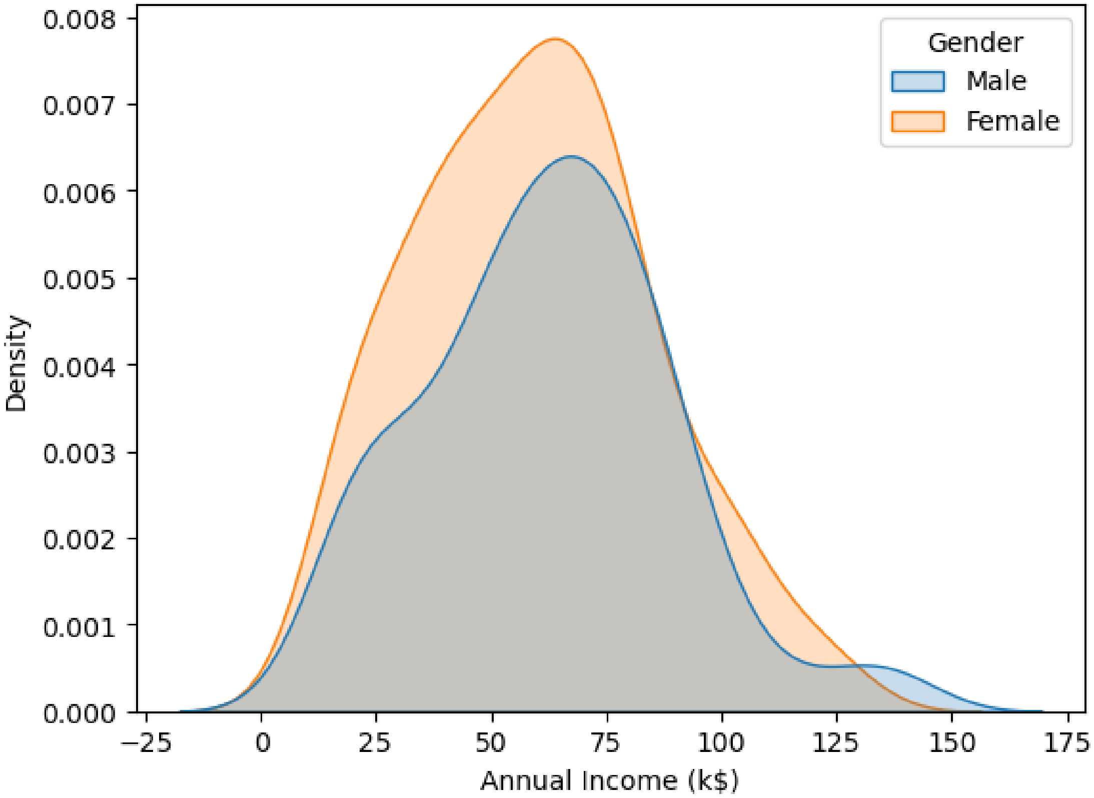
```
[9]: sns.kdeplot(data=df, x='Annual Income (k$)', hue='Gender', shade=True);
```

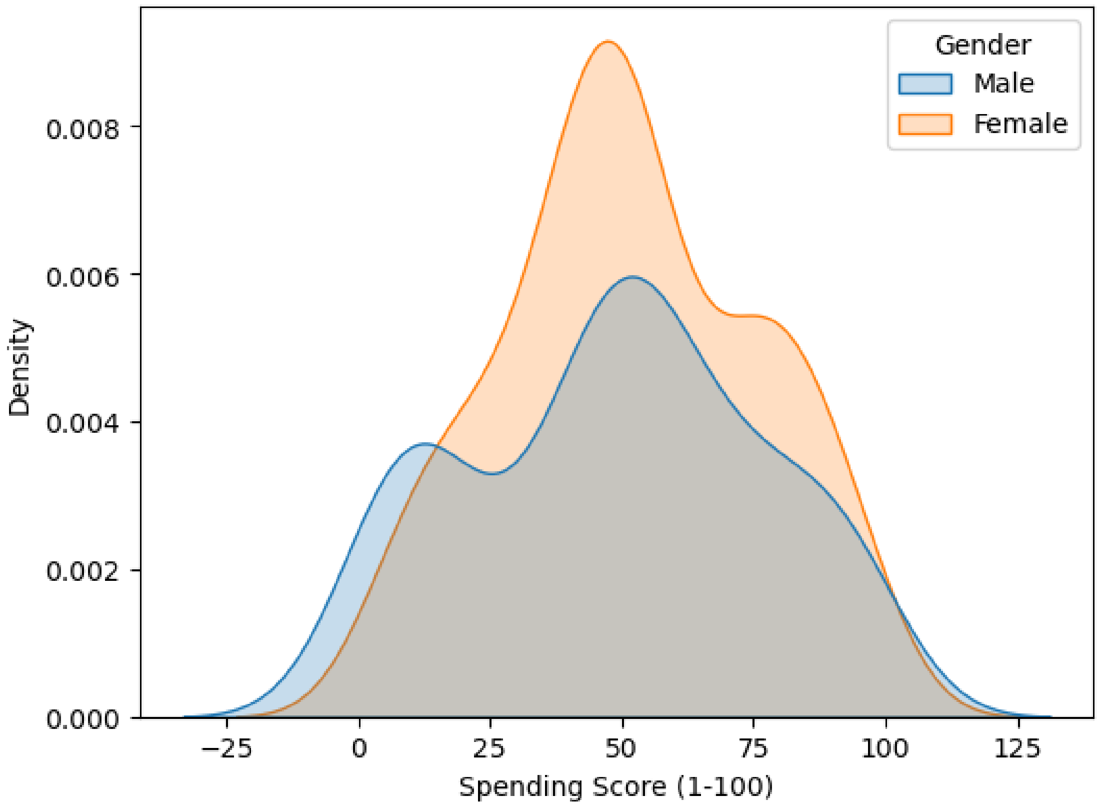


[10]:

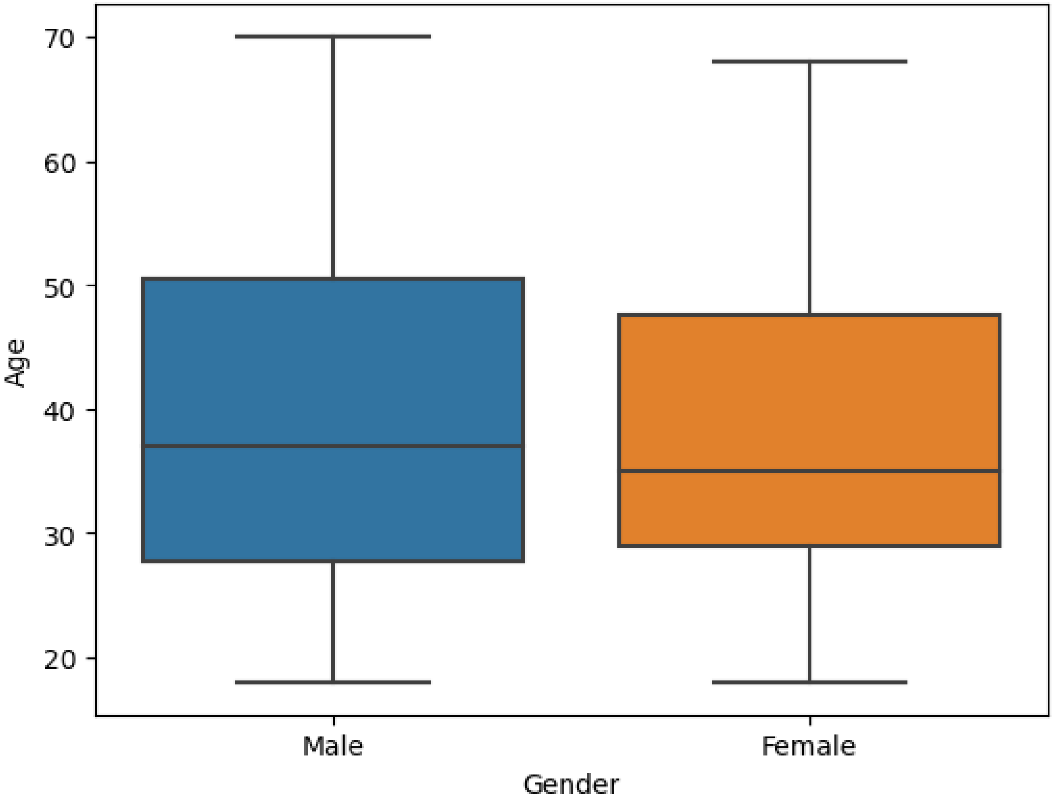
```
columns = ['Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.kdeplot(data=df,x=i, hue='Gender', shade=True);
```

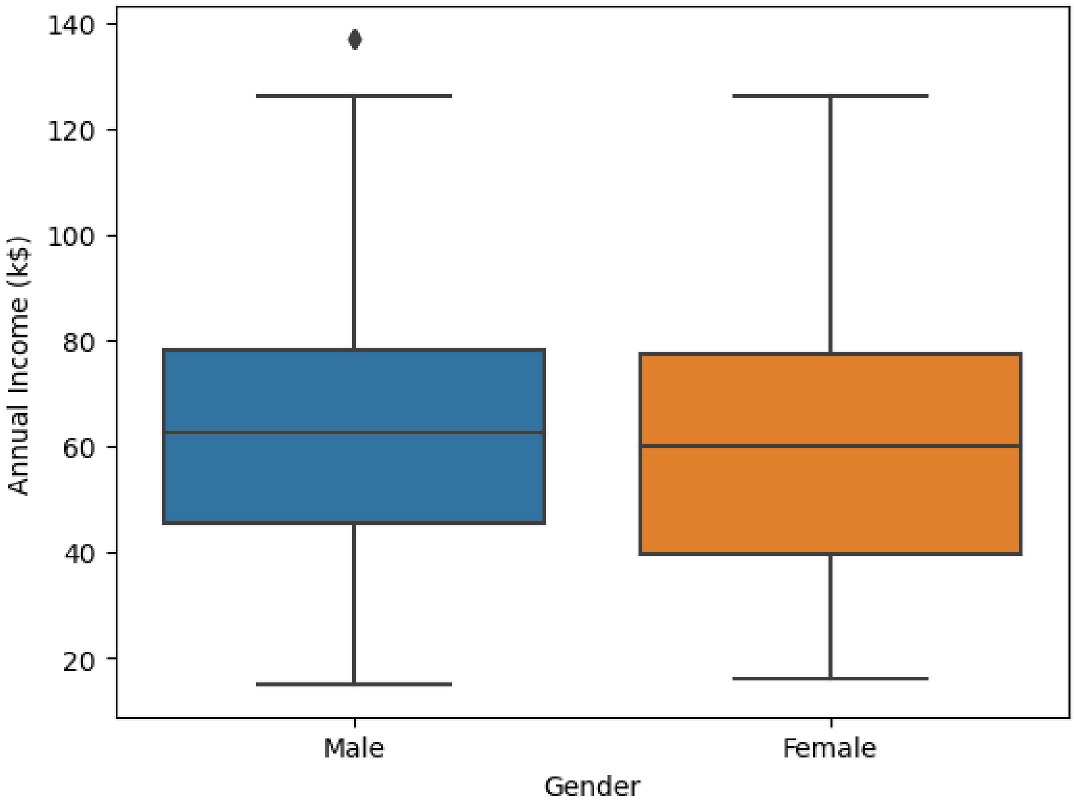

0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Density
0
20
40
60
80
Age
Gender
Male
Female

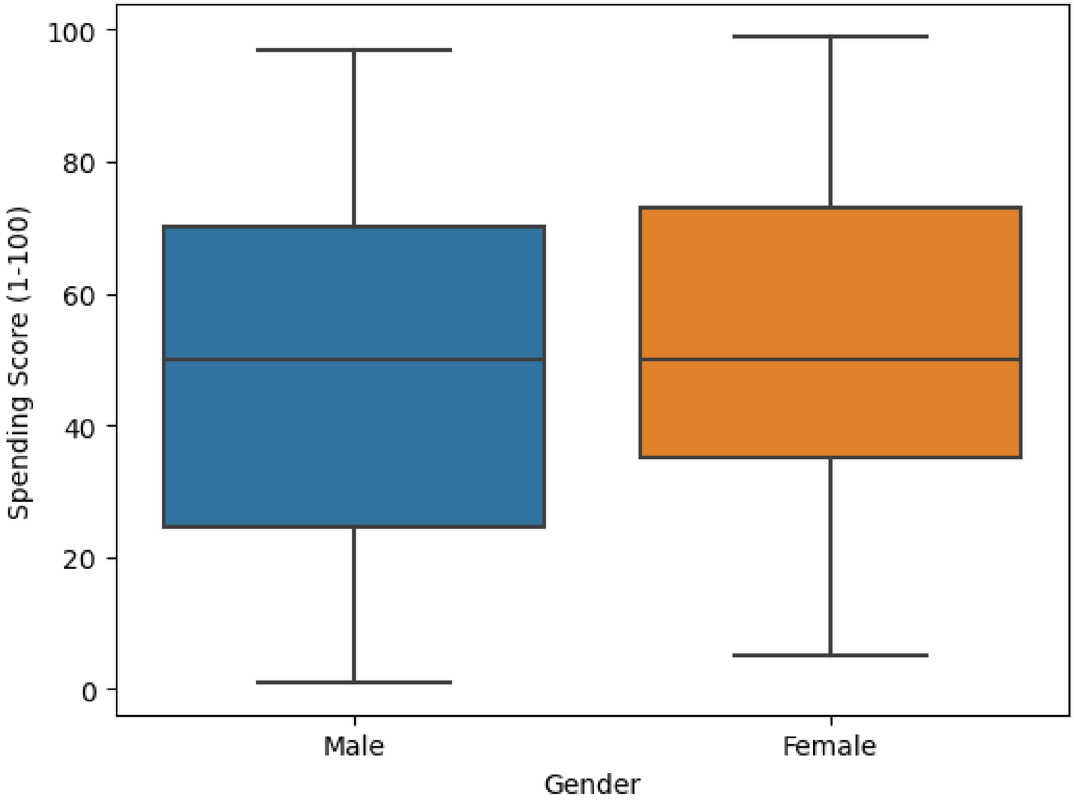

Gender
Male
Female
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
−25
0
25
50
75
100
125
150
175
Annual Income (k$)



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


70
60
50
40
30
20
Age
Male
Female
Gender

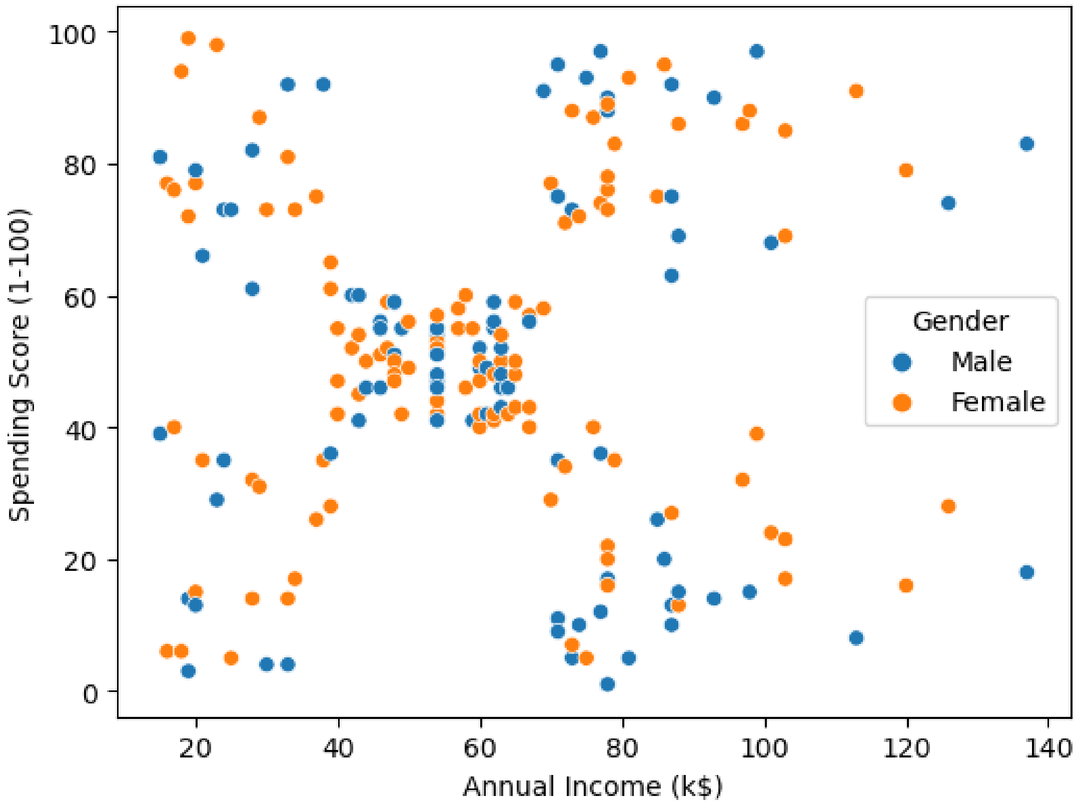

140
120
100
80
60
40
20
Annual Income (k$)
Male
Female
Gender



2 Bivariate Analysis

[12]:
```
#import seaborn as sns

sns.scatterplot(data=df, x='Annual Income (k$)', y='Spending Score (1-100)',␣
 ↪hue='Gender');
```



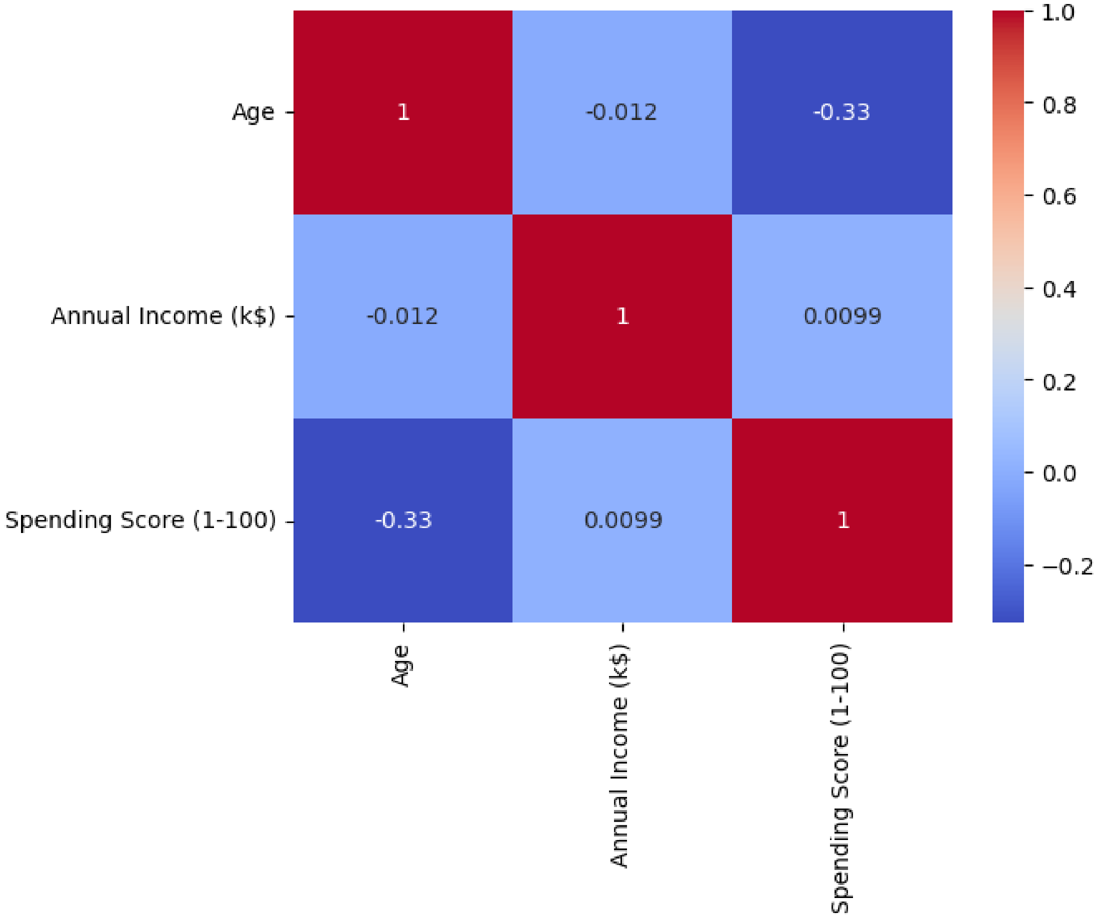
[13]:
```
df.groupby('Gender')[['Age', 'Annual Income (k$)', 'Spending Score (1-100)']].
    ↪mean()
```

[13]:
```
              Age  Annual Income (k$)  Spending Score (1-100)
Gender
Female  38.098214           59.250000               51.526786
Male    39.806818           62.227273               48.511364
```

[14]:
```
corr_matrix = df.corr(numeric_only=True)
corr_matrix
```

[14]:
```
                             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
```

[15]:
```
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm');
```



[16]:

```
income_by_gender = df.groupby('Gender')['Annual Income (k$)'].agg(['mean',␣
 ↪'sum', 'count'])
print(income_by_gender)
```

```
             mean   sum  count
Gender
Female  59.250000  6636    112
Male    62.227273  5476     88
```

[17]:

```
avg_age = df['Age'].mean()
avg_age
```

[17]: 38.85

[18]:

```
avg_income = df['Annual Income (k$)'].mean()

avg_income
```

[18]:
```
60.56
```

[19]:
```
df['Gender'].value_counts()
```

[19]:
```
Gender
Female    112
Male       88
Name: count, dtype: int64
```

[20]:
```
df['Gender'].value_counts(normalize=True)
```

[20]:
```
Gender
Female    0.56
Male      0.44
Name: proportion, dtype: float64
```

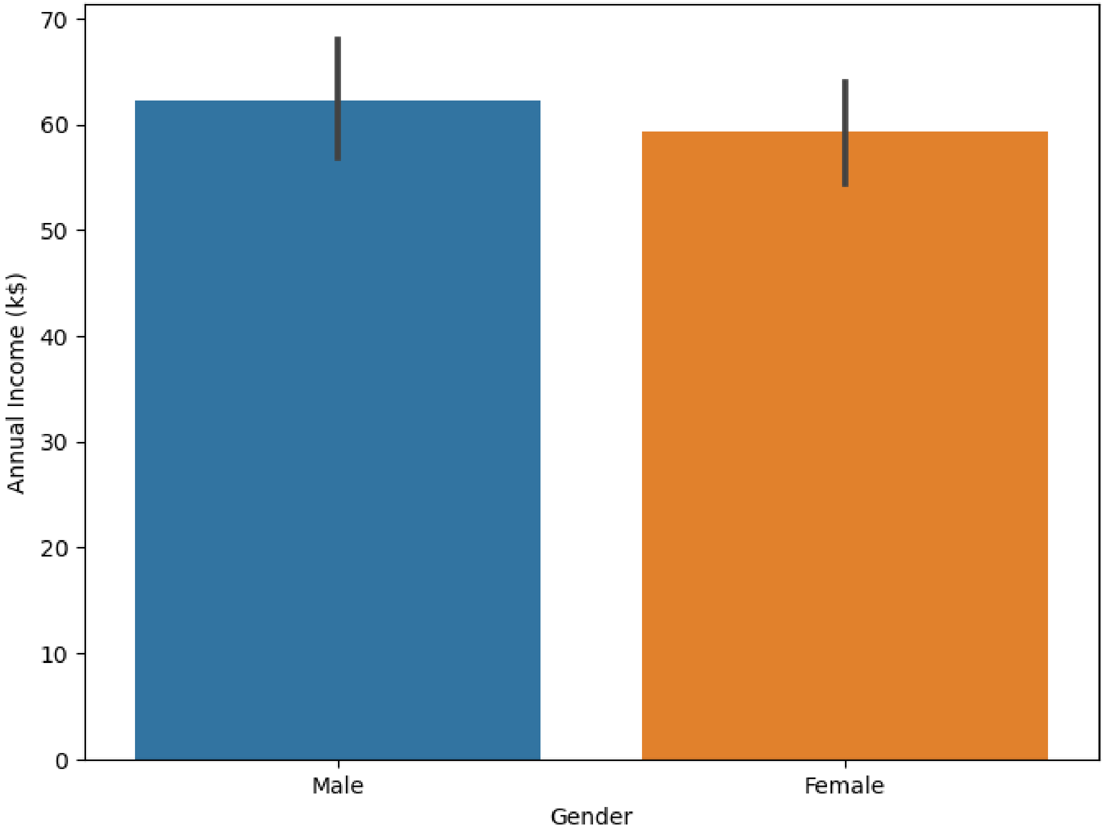
[21]:
```
df['Gender'].value_counts(normalize=True).apply(lambda x: f'{x*100:.0f}%')
```

[21]:
```
Gender
Female    56%
Male      44%
Name: proportion, dtype: object
```

[22]:
```
df.columns
```

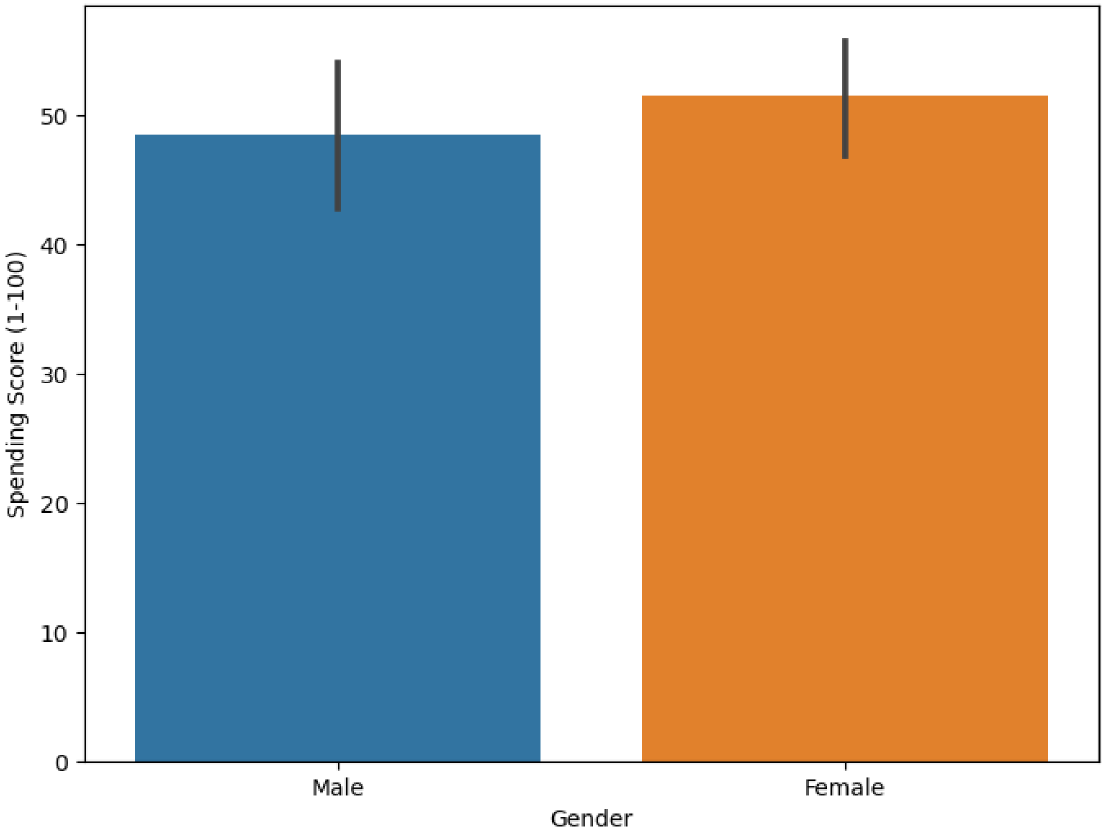
[22]:
```
Index(['Gender', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)'],
dtype='object')
```

[23]:
```
# Plotting
plt.figure(figsize=(8, 6))
sns.barplot(data=df, x='Gender', y='Annual Income (k$)');
```



[24]:

```
# Plotting
plt.figure(figsize=(8, 6))
sns.barplot(data=df, x='Gender', y='Spending Score (1-100)');
```



50
40
30
20
10
0
Spending Score (1-100)
Male
Female
Gender

[ ]:

[ ]: