Objective:

Customer segmentation is a technique in which we divide the customers based on their purchase history, gender, age, interest, etc. It is useful to get this information so that the store can get help in personalizing marketing and provide customers with relevant deals.

In [1]:

```
#importing librabries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Explanation:

The data set that has been provided is on mall customers. Customer segmentation is helpful in understanding the insights of the data. Also, with the help of data visualization we can analyze the data more accurately and efficiently.

In [2]:

```
#importing dataset
df=pd.read_csv('C:/salandri-nirusha-data _science/customer/Mall_Customers.csv')
```

In [3]:

```
df
```

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
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

[describe()] function tries to give the statistical view about our data.

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]:

```
#Head call returns the top 5 rows from the data
df.head()
```

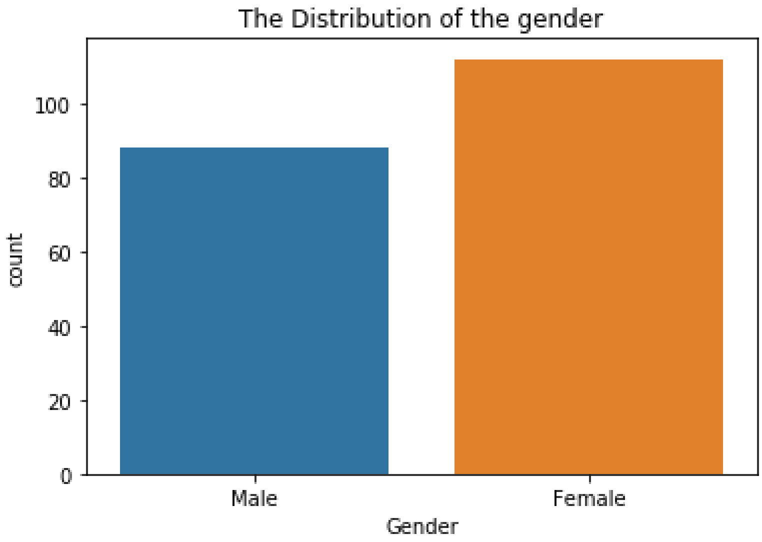
Out[5]:

	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

TO understand the distribution of the categorical variables throughout the data set, a simple count plot is uded as given below:

In [6]:

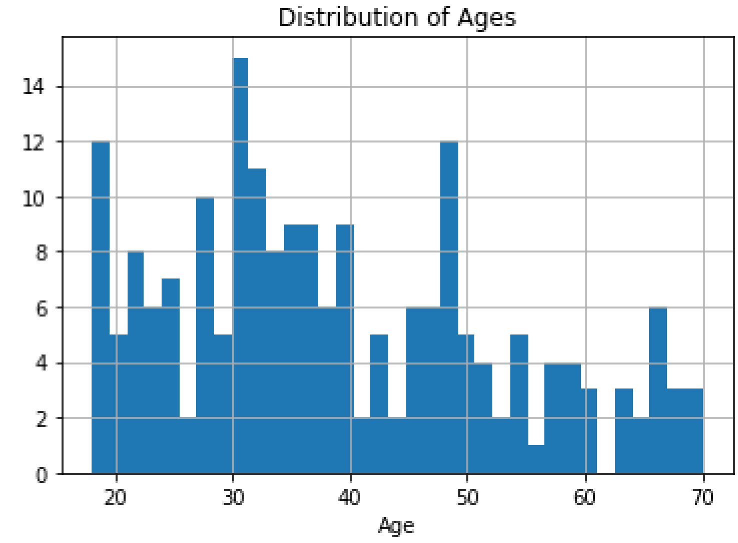
```
#see the distribution of gender
sns.countplot(x='Gender',data=df);
plt.title("The Distribution of the gender");
```



As we can see from the above graph the female customers are more in number than male and let's understand the Age distribution using histogram. The plt.hist() function creates histogram plots.bins denotes the number of bins on the histogram.

In [7]:

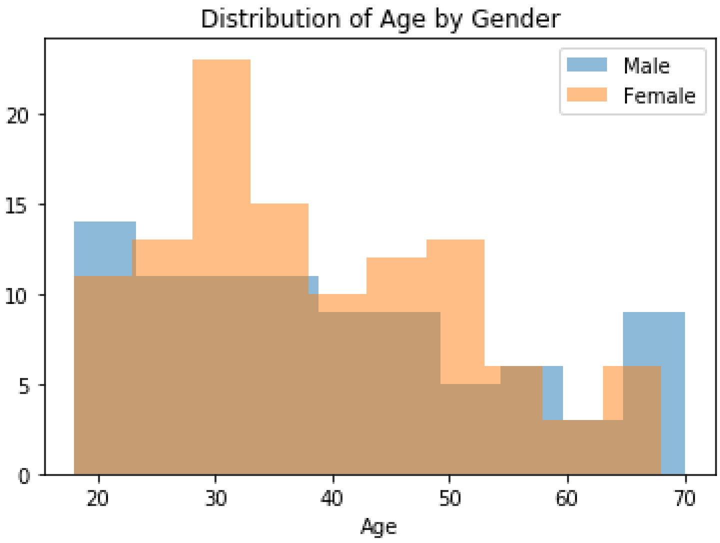
```
# Creating a histogram to understand the distribution of Ages
df.hist('Age',bins=35);
plt.title('Distribution of Ages');
plt.xlabel('Age');
```



From the above figure we can depict that the ages are mostly between 30 and 40 .If we recall the describe() call results, the average age was 38.The distribution is a right-skewed.

In [8]:

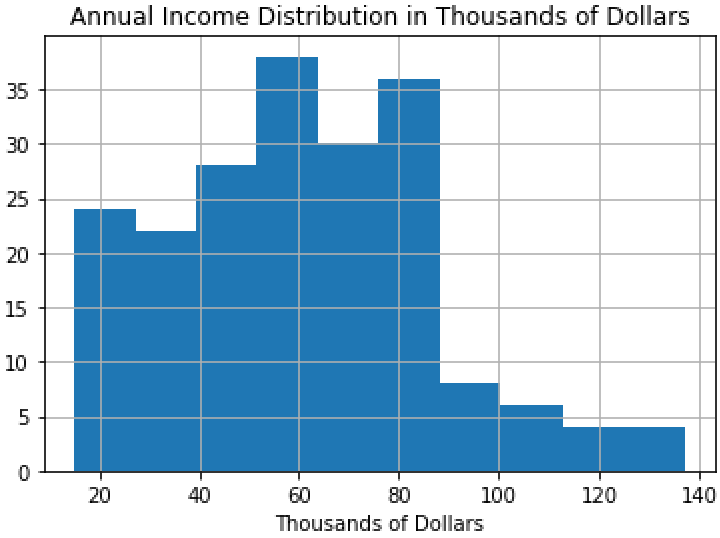
```
#To see the distribution of age by gender
plt.hist('Age',data=df[df['Gender']=='Male'],alpha=0.5,label='Male');
plt.hist('Age',data=df[df['Gender']=='Female'],alpha=0.5,label='Female');
plt.title('Distribution of Age by Gender');
plt.xlabel('Age');
plt.legend();
```



Setting the opacity(alpha) and Plot legends give meaning to a visualization, assigning meaning to the various plot elements. The men in this data set tend to be more of young age that in women.you can see the spike around the age of 30-35 for the women is where the majority of them fall.There are also more middle-aged women in this dataset than men .There is a significant amount in the senior men list[65-70] the last bucket.

In [9]:

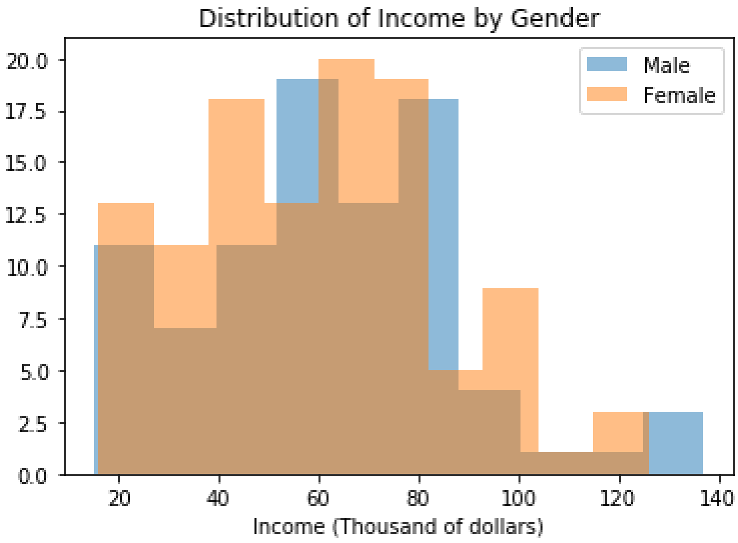
```
#To see the distribution of income
df.hist('Annual Income (k$)');
plt.title('Annual Income Distribution in Thousands of Dollars');
plt.xlabel('Thousands of Dollars');
```



Much of the income lie between the 55,000 -85,000 dollar buckets.Do you think gender can impact on this or not???

In [10]:

```
#Histogram of income by Gender
plt.hist('Annual Income (k$)',data=df[df['Gender']=='Male'],alpha=0.5,label='Male');
plt.hist('Annual Income (k$)',data=df[df['Gender']=='Female'],alpha=0.5,label='Female'
);
plt.title('Distribution of Income by Gender');
plt.xlabel('Income (Thousand of dollars)');
plt.legend();
```



In [11]:

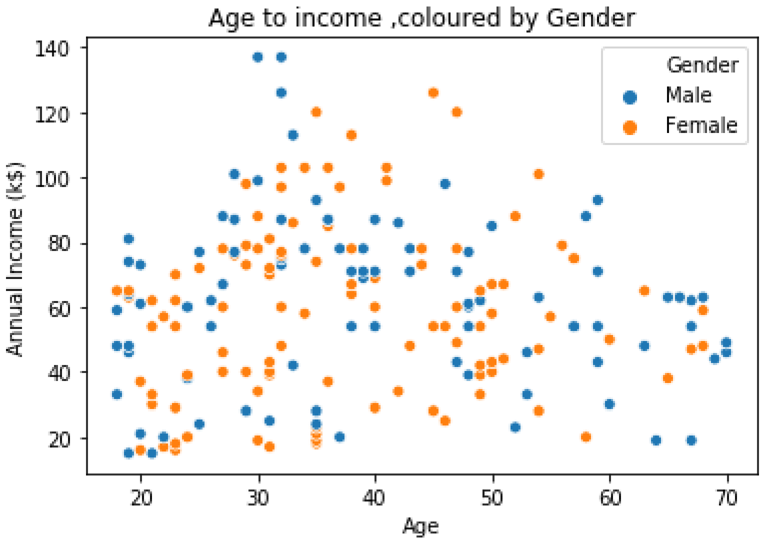
```
#create Data sets by Gender to save time in the furture since Gender seems to significa
ntly impact other Variable
male_customers=df[df['Gender']=='Male']
Female_customers=df[df['Gender']=='Female']

#Print the average spending score for men and women
print(male_customers['Spending Score (1-100)'].mean())
print(Female_customers['Spending Score (1-100)'].mean())
```

```
48.51136363636363
51.526785714285715
```

From the above two codes [9,10] It is observed that women make less money than the men,if we go by this data set it'll lead to the question of how their spending score looks like and how it compares?? If we look at the avg spending score of men=48.5 and avg spending score of female= 51.5 Hands down :),though women earn less they just love to spend more money in the Malls.

In [12]:

```
sns.scatterplot('Age','Annual Income (k$)',hue='Gender',data=df);
plt.title('Age to income ,coloured by Gender');
```



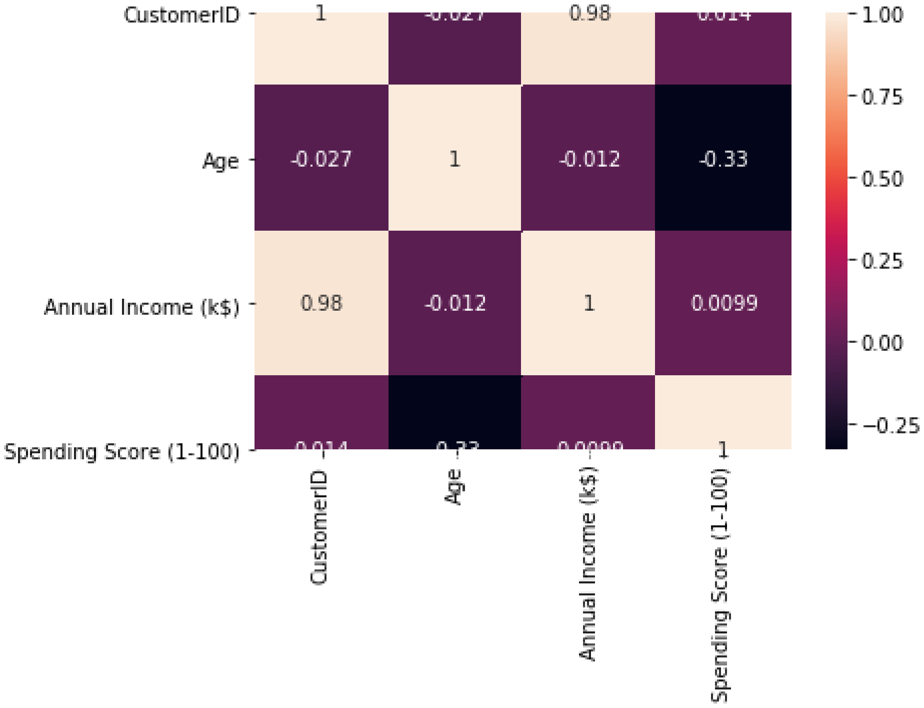
Correlation Heat map of each variable :

In [13]:

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

Out[13]:

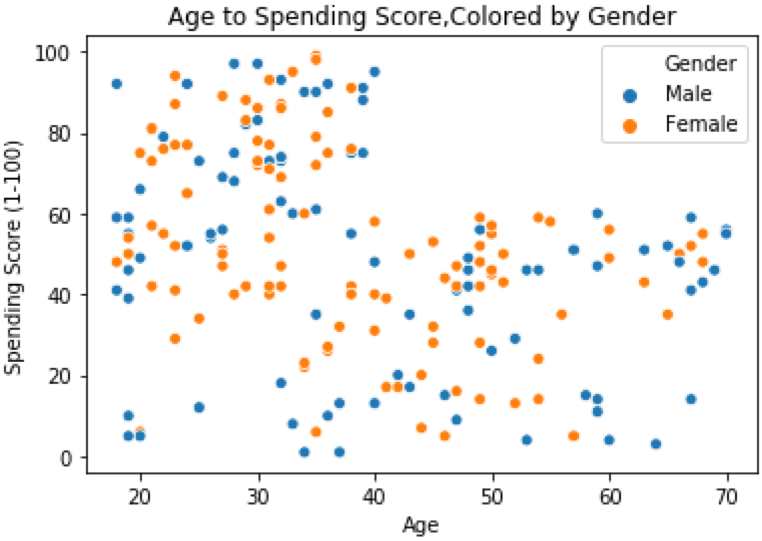
```
<matplotlib.axes._subplots.AxesSubplot at 0x184cbe198c8>
```



when you look at the black boxes from the graph, It's a negative correlation so the older a customer is in the data set, the lower their spending score.but because it's 0.33,it's not a strong correlation at all. let's look into this point in a clear view.

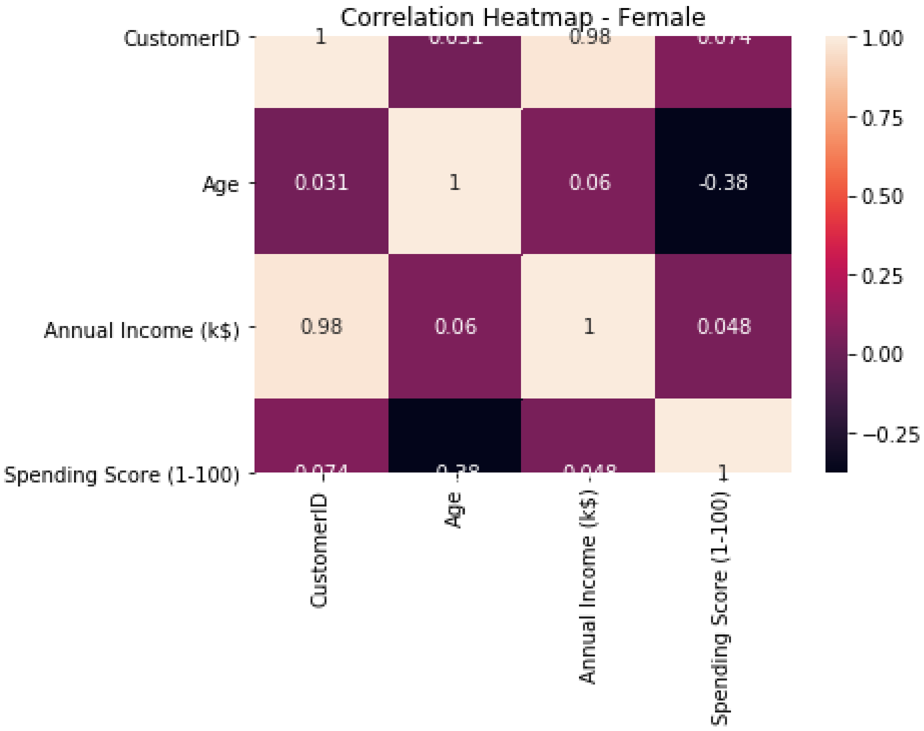
In [14]:

```
sns.scatterplot('Age','Spending Score (1-100)',hue='Gender',data=df);
plt.title('Age to Spending Score,Colored by Gender');
```



In [15]:

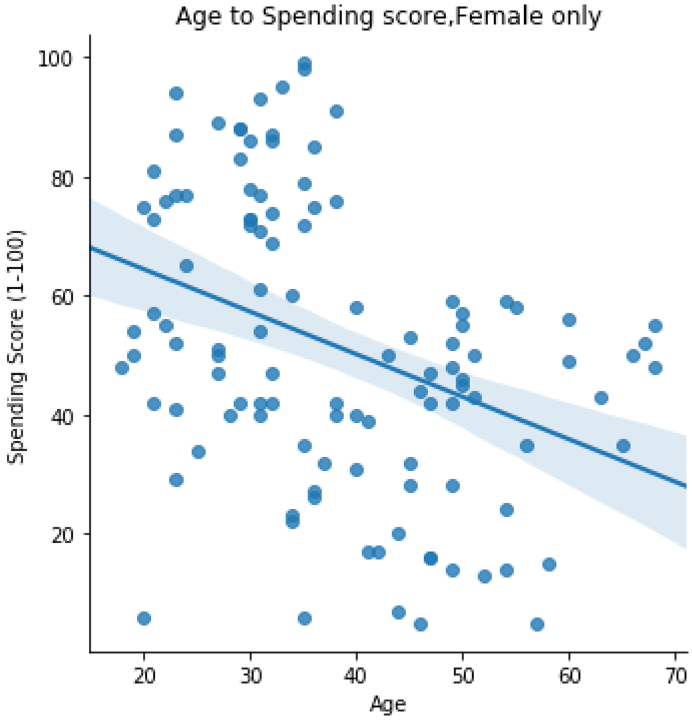
```
sns.heatmap(Female_customers.corr(),annot=True);
plt.title('Correlation Heatmap - Female');
```



Age more strongly affects spending scores for women in this case .Now we can Zoom in on the women's spending score to age relationship.

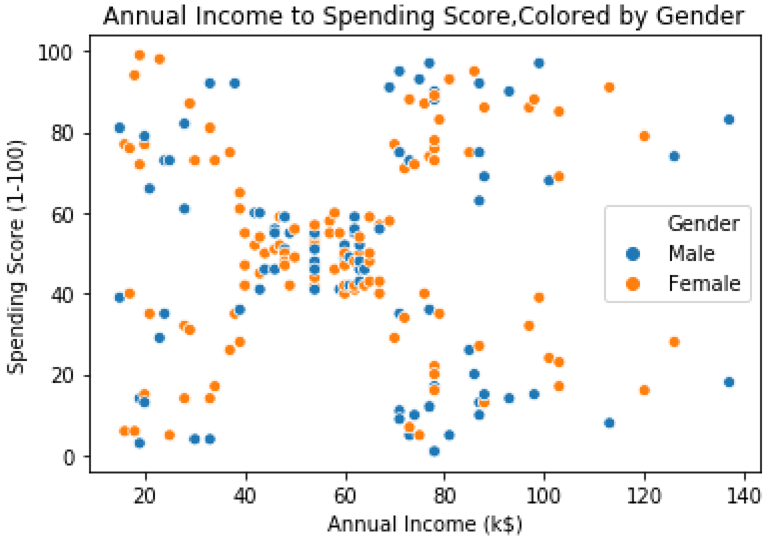
In [16]:

```
sns.lmplot('Age','Spending Score (1-100)',data=Female_customers);
plt.title('Age to Spending score,Female only');
```



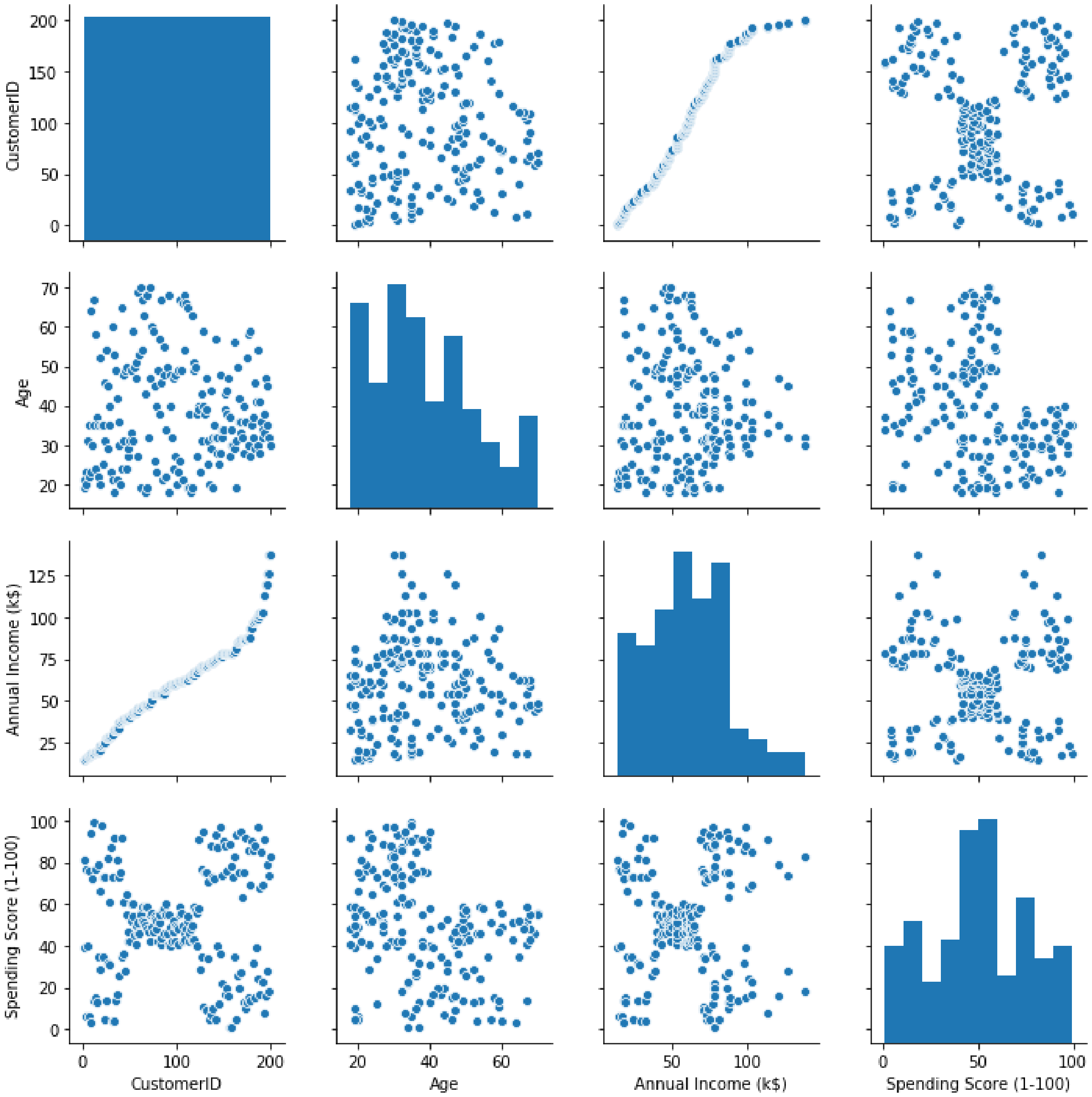
In [17]:

```
sns.scatterplot('Annual Income (k$)','Spending Score (1-100)',hue='Gender',data=df);
plt.title('Annual Income to Spending Score,Colored by Gender');
```



In [18]:

```
sns.pairplot(df);
```



There are some patterns here.Though the correlation is not high and accurate.It would be helpful to plan on how to gather more data set that has more features.The more features,the better understanding of the data and insights.To make a point here,a good data scientist has a curiosity to dive into the data just by figuring out with the smallest of the data and give you the simplest point which can tale the model.

In [ ]: