Project: Mall customers segmentation using K Means clustering Algorithm in python

About Project: Using a data set of mall customers, used a K Means clustering algorithm to learn about the clustering groups.

Data Source: Mall customers segmentation.csv

Otuekong Nyong • 17.03.2023

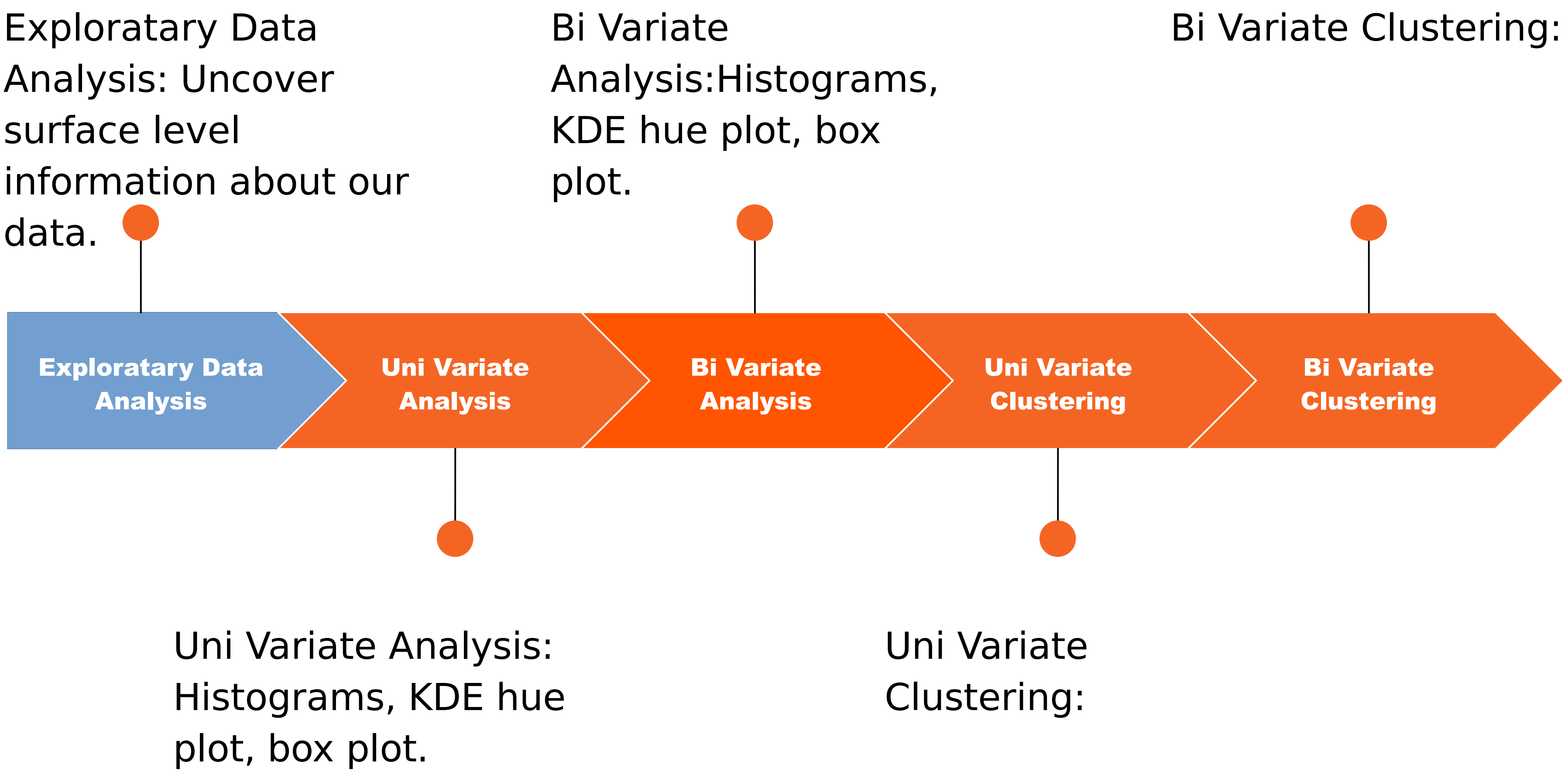
Overview

Steps To Completion

- Perfomrm some exploratary data analysis(UniVariate, BiVariate)
- Use Kmeans clustering algorithm to do UniVariate and BiVariate clustering.
- Perform data analysis

Tools Used

Microsoft Excel



Exploratary Data Analysis: Uncover surface level information about our data.
Bi Variate Analysis:Histograms, KDE hue plot, box plot.
Bi Variate Clustering:
Exploratary Data Analysis
Uni Variate Analysis
Bi Variate Analysis
Uni Variate Clustering
Bi Variate Clustering
Uni Variate Analysis: Histograms, KDE hue plot, box plot.
Uni Variate Clustering:

Exploratary Data Analysis

Exploratary Data Analysis

Uni Variate Analysis

- Create a histogram to showcase the annual income and density distribution using seaborn distplot
- Visualize the split annual income with only kde using the hue parameter
- Used a for loop too see how gender compares with age annual income and spending score with a kde hue plot and a box plot
- Using value count we count the values of the column ‘Gender’ to see how many are there
- Perform analysis

Exploratary Data Analysis

Bi Variate Analysis

Bivariate analysis works with two variables

- Use a scatterplot to create a Bicluster, graphing annual income over spending score
- using groupby to see mean values by gender perform analysis
- Create a heat map, and for parameters we can use annotations as a parameter, and also use c map which is the color mapping using cool and warm
- Perform final analysis

Univariate Analysis

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

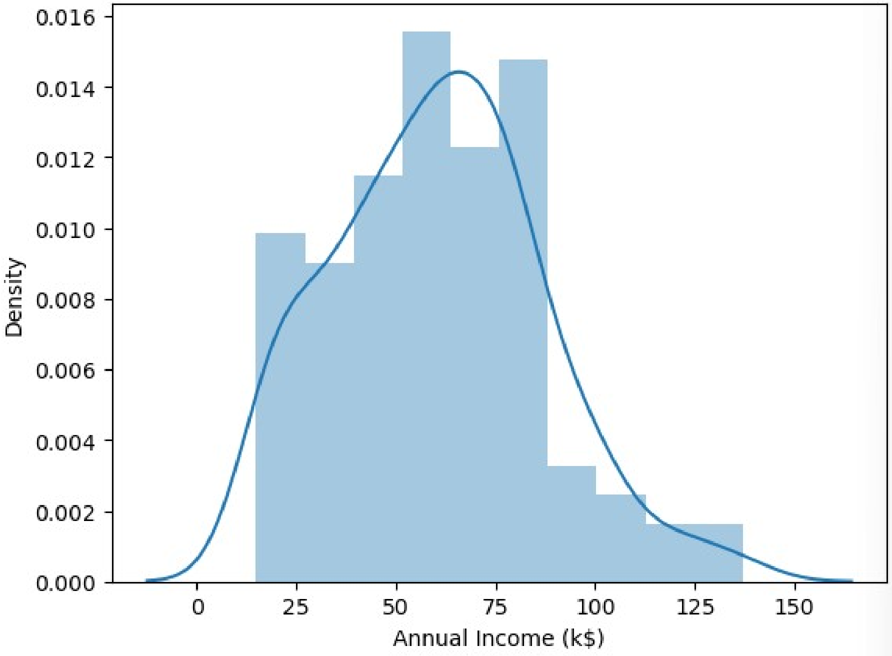
Out[4]:

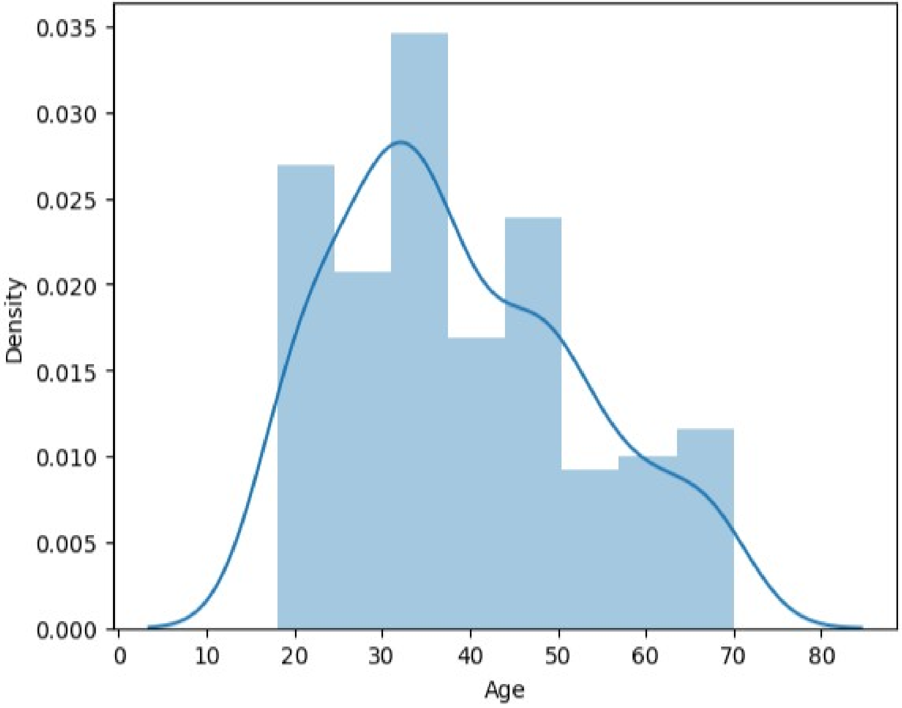
	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Create a histogram to showcase the annual income and density distribution using seaborn distplot

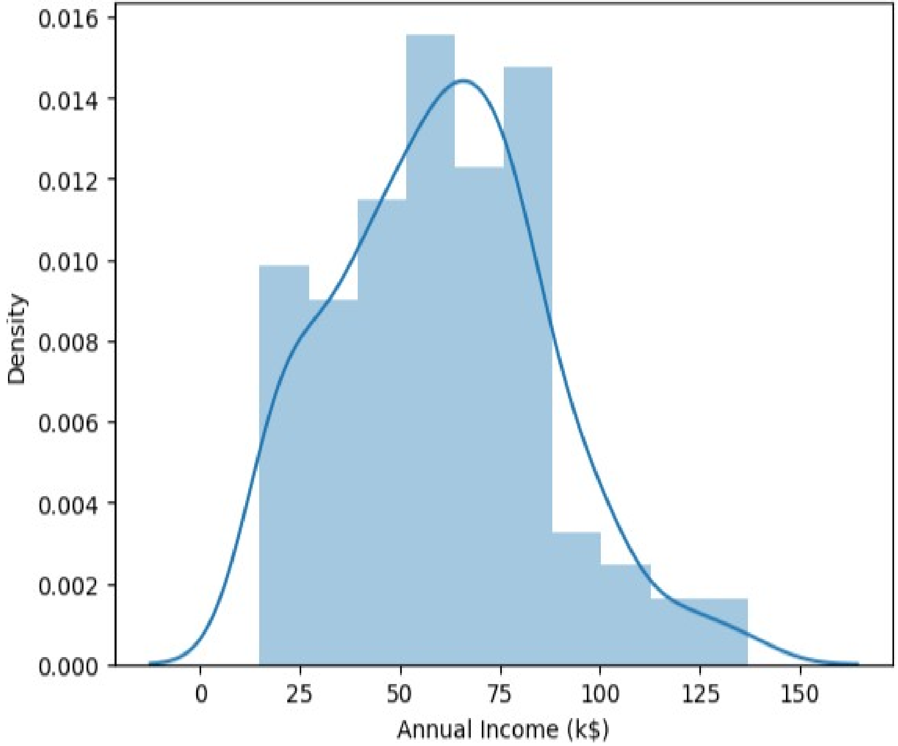
Uni Variate

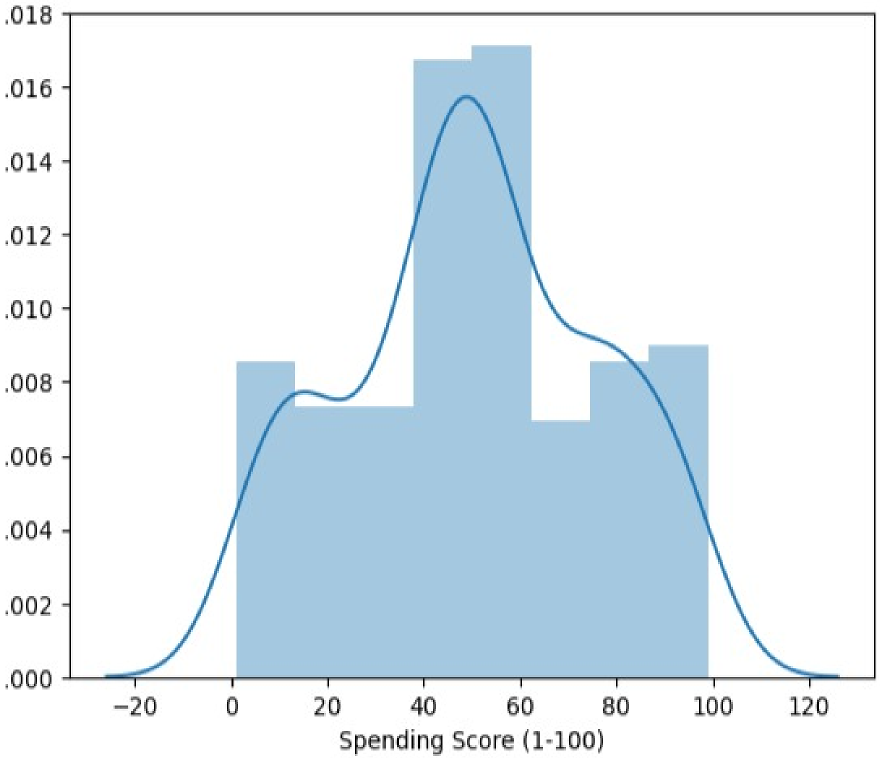
<AxesSubplot:xlabel='Annual Income (k$)', ylabel='Density'>



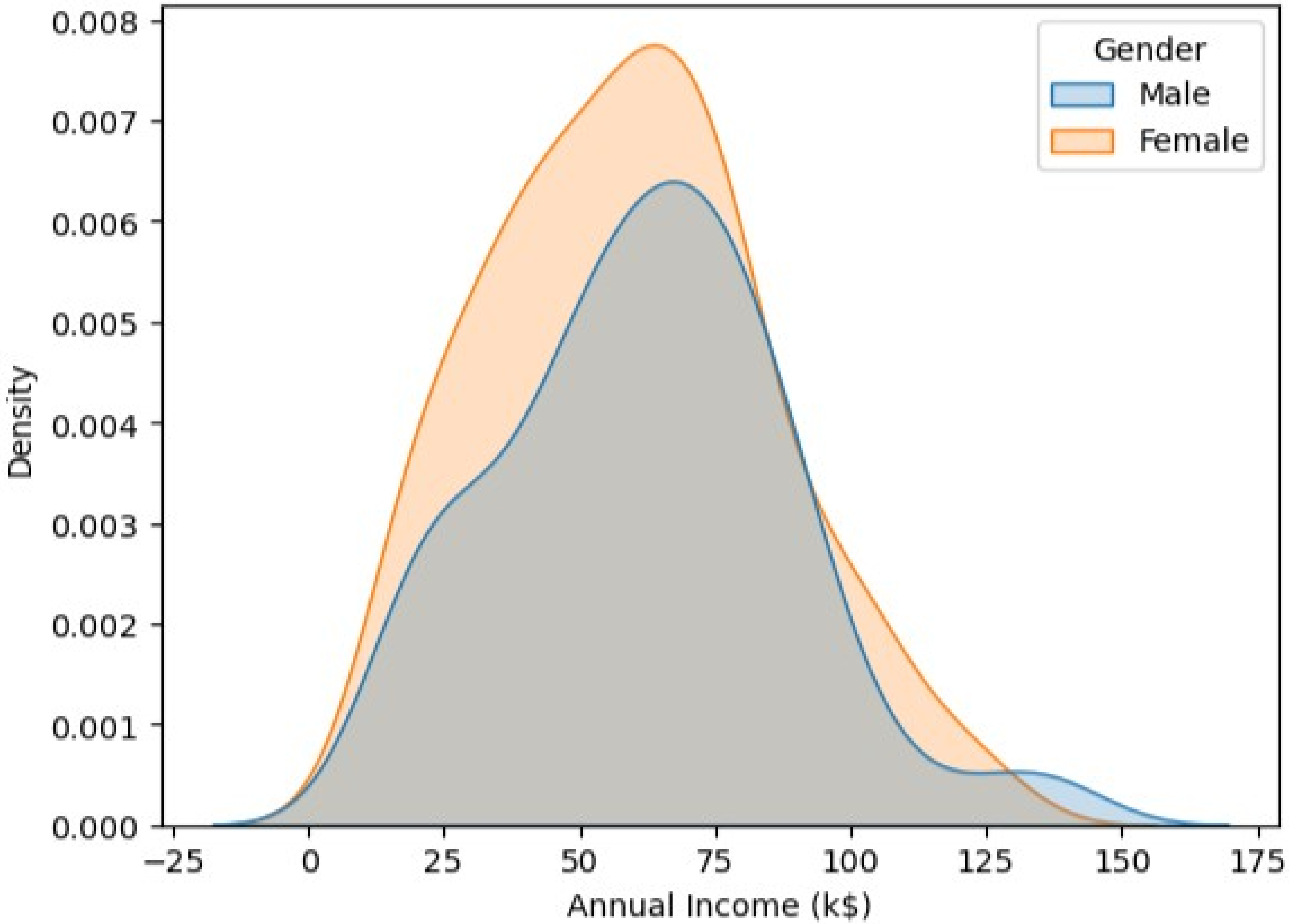


Uni Variate


Density
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
25
50
75
100
125
150
Annual Income (k$)

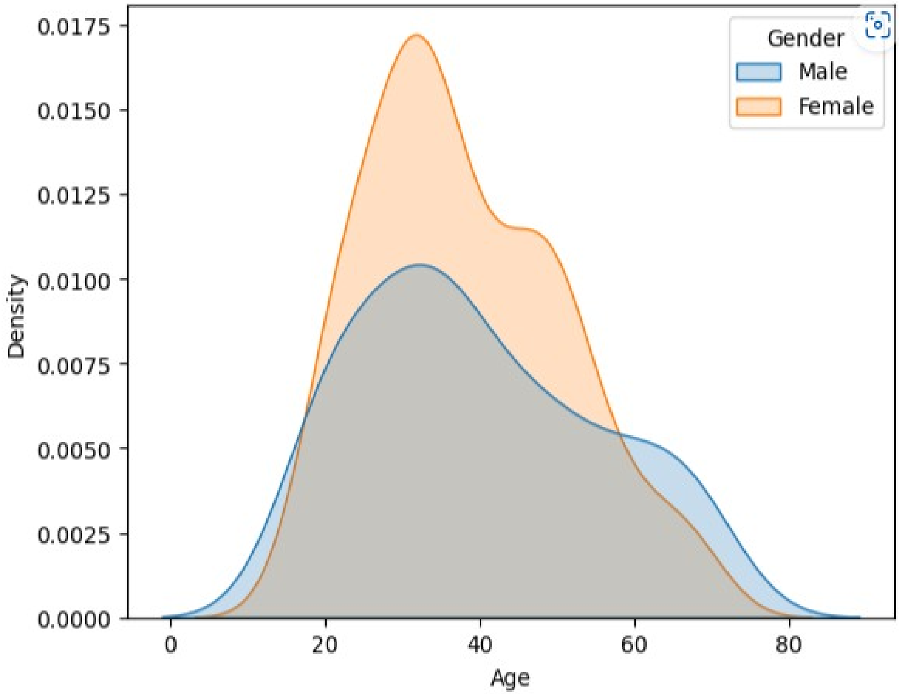

.018
.016
.014
.012
.010
.008
.006
.004
.002
.000
−20
0
20
40
60
80
100
120
Spending Score (1-100)

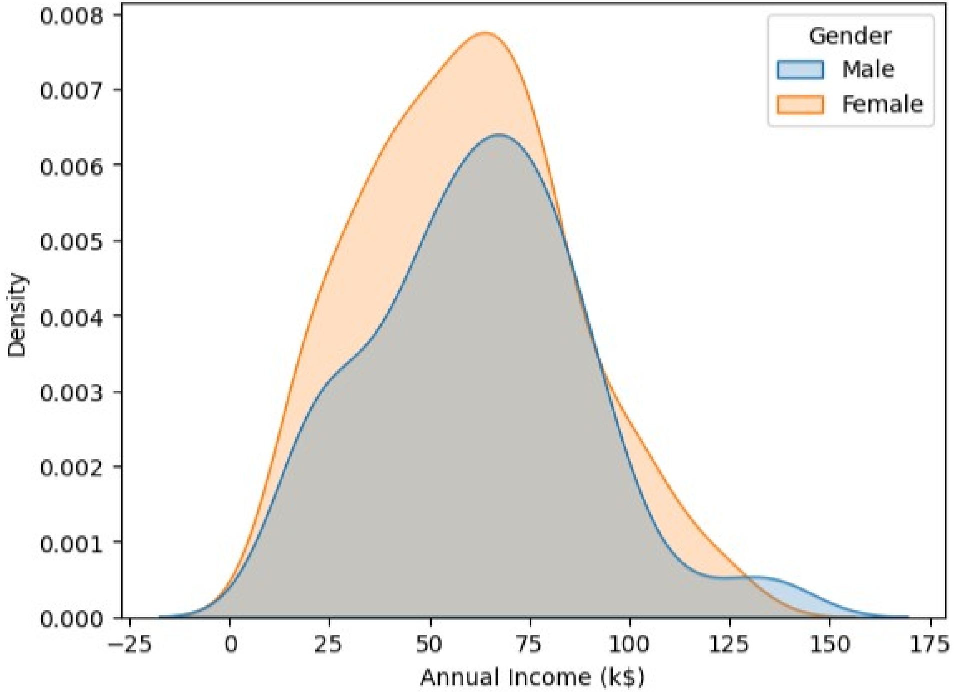
```
sns.kdeplot(mk['Annual Income (k$)'],shade=True,hue=mk['Gender']);
```



Visualize the split annual income with only kde using the hue parameter

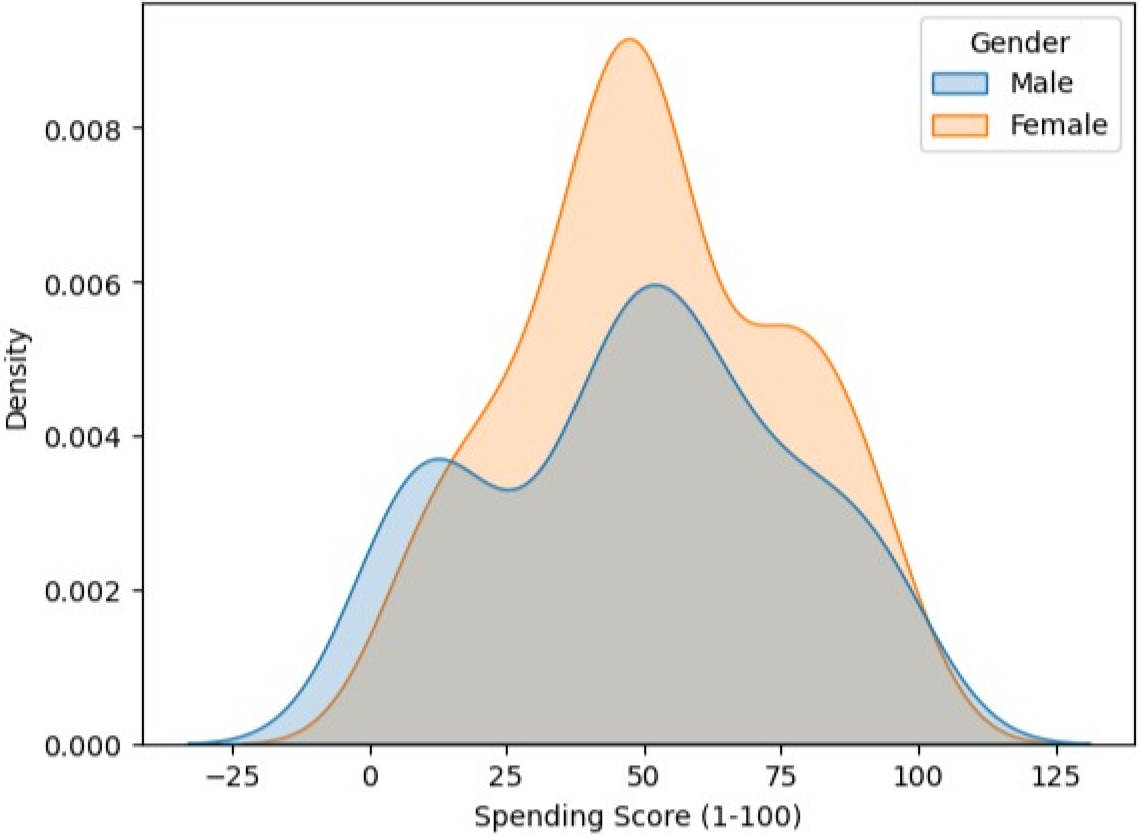
Uni Variate

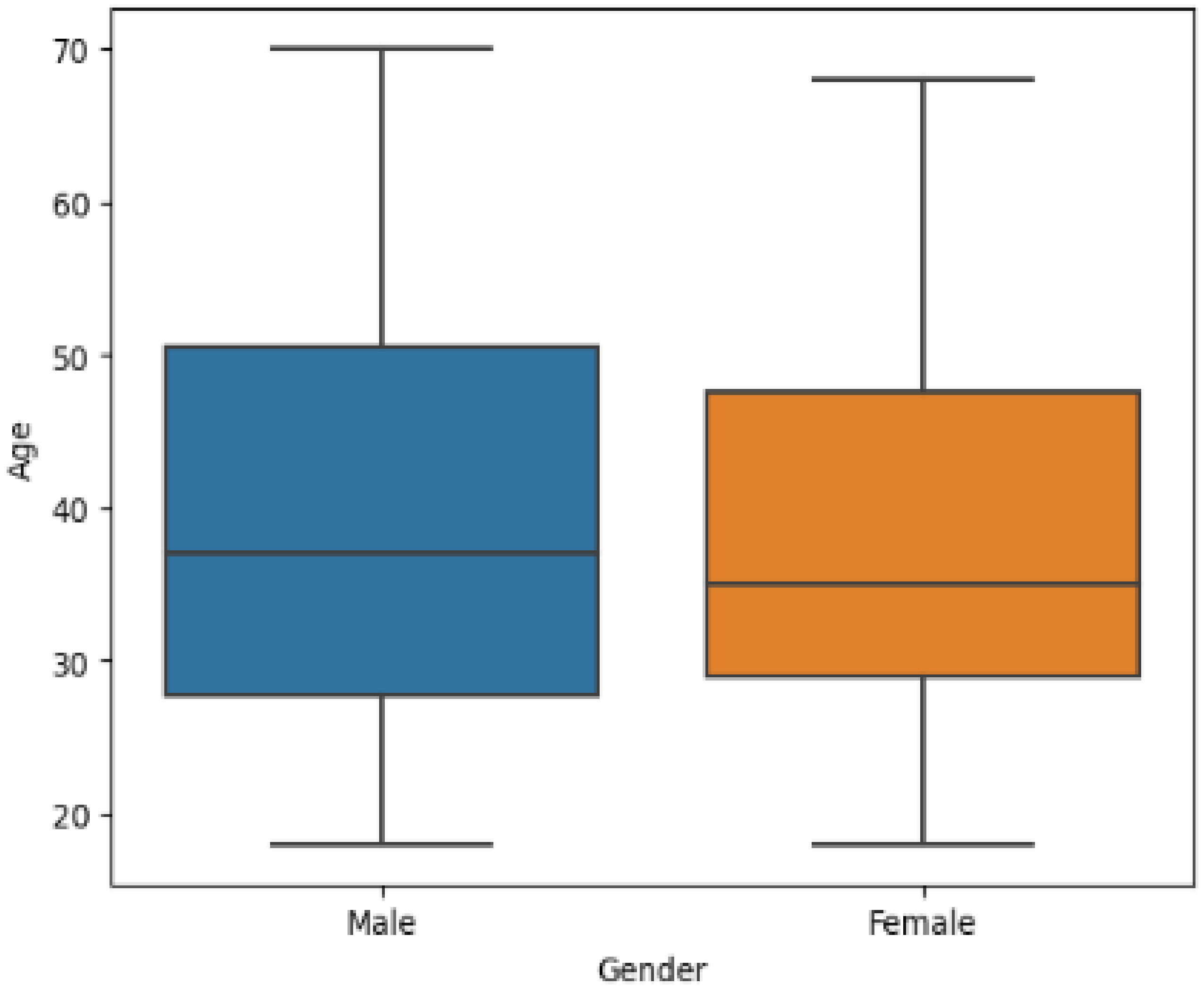

Gender
Male
Female
Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
20
40
60
80
Age


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

Uni Variate

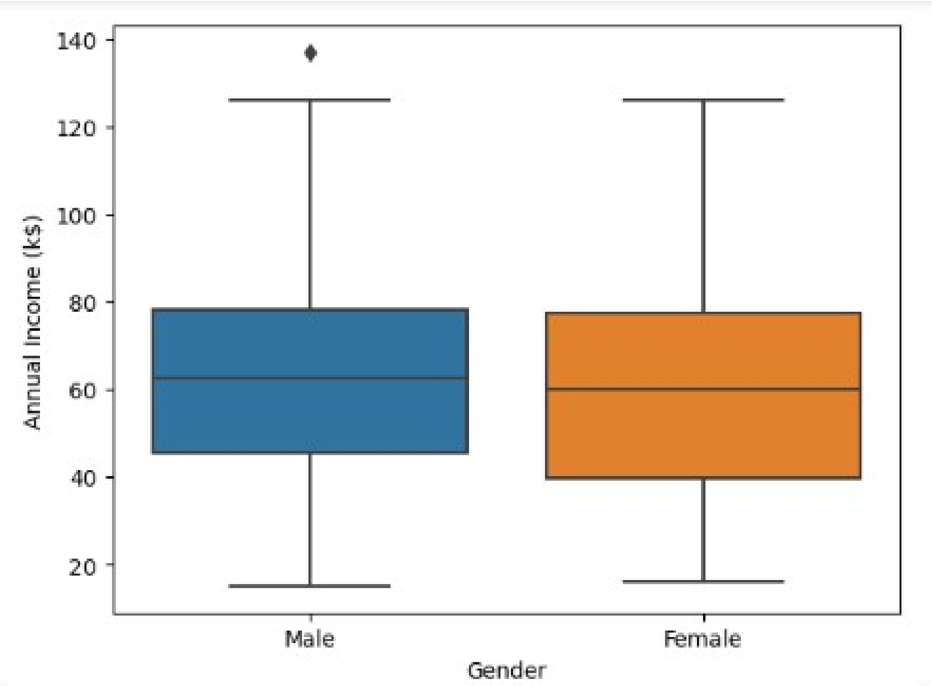
Used a for loop too see how gender compares with age annual income and spending score in a kde hue plot



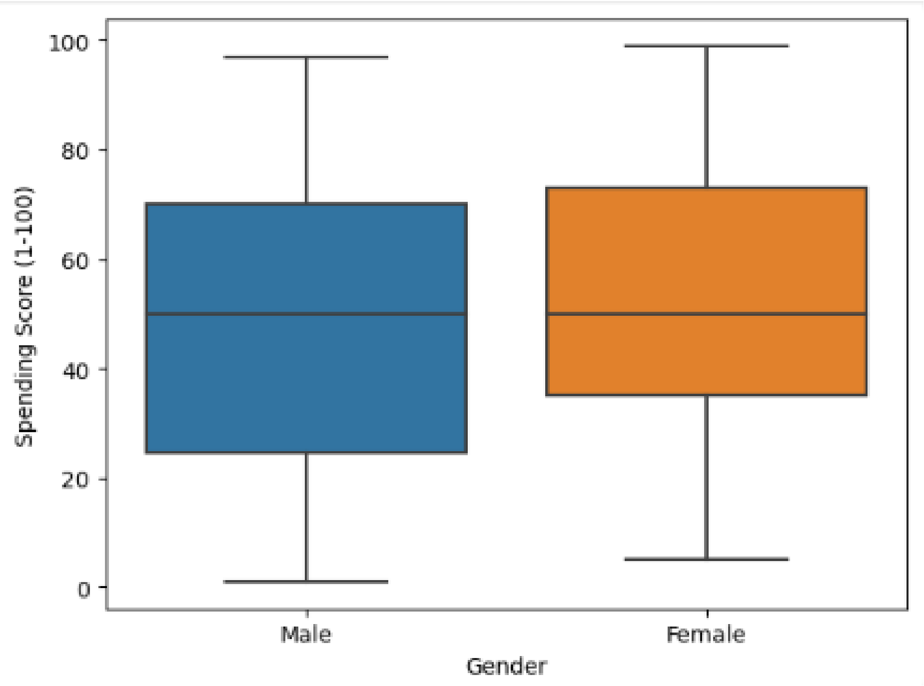


Used a for loop too see how gender compares with age annual income and spending score with a box plot

Uni Variate



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



100
80
60
40
20
0
Spending Score (1-100)
Male
Female
Gender

Uni Variate

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

```
Out[21]:

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

- **so through this we’ve discovered that 56% of customers are female and 44% percent are male**

Progress – Uni Variate Analysis

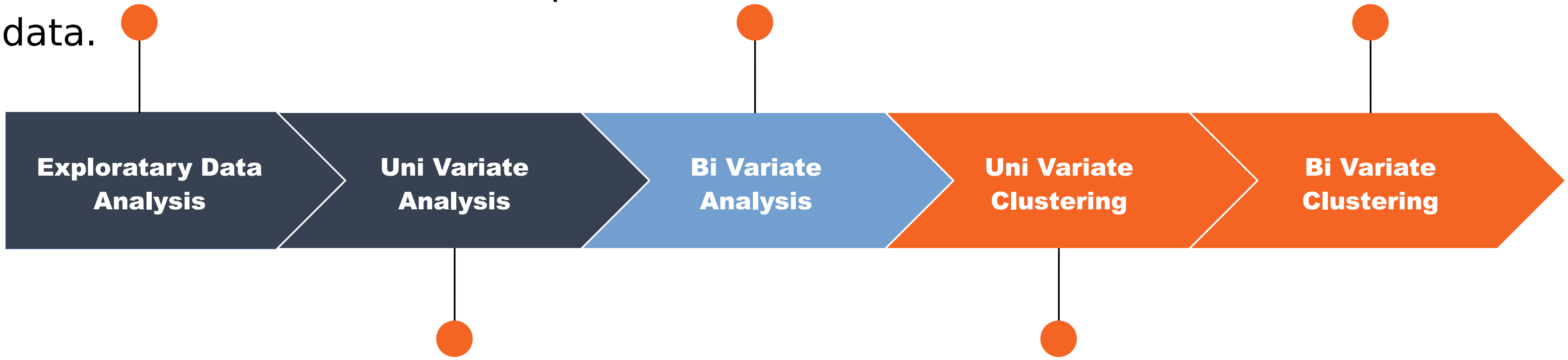
Analysis

- Used Univariate analysis to explore our data and turned it into information we understand.
- From the Histogram we can see the spending score is isolated 40-50
- Using value count we count the values of the column ‘Gender’ to see the gender values. Through this we’ve discovered that 56% of customers are female and 44% percent are male

Exploratary Data Analysis: Uncover surface level information about our data.

Bi Variate Analysis:Histograms, KDE hue plot, box plot.

Bi Variate Clustering:



Uni Variate Analysis: Histograms, KDE hue plot, box plot.

Uni Variate Clustering:

Bi Variate Analysis

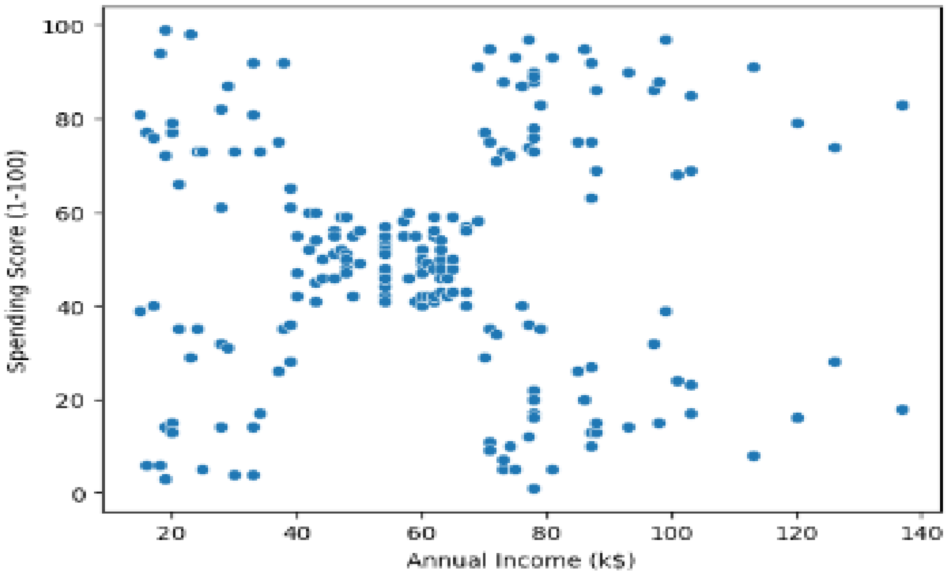
Bivariate Analysis

In [22]:

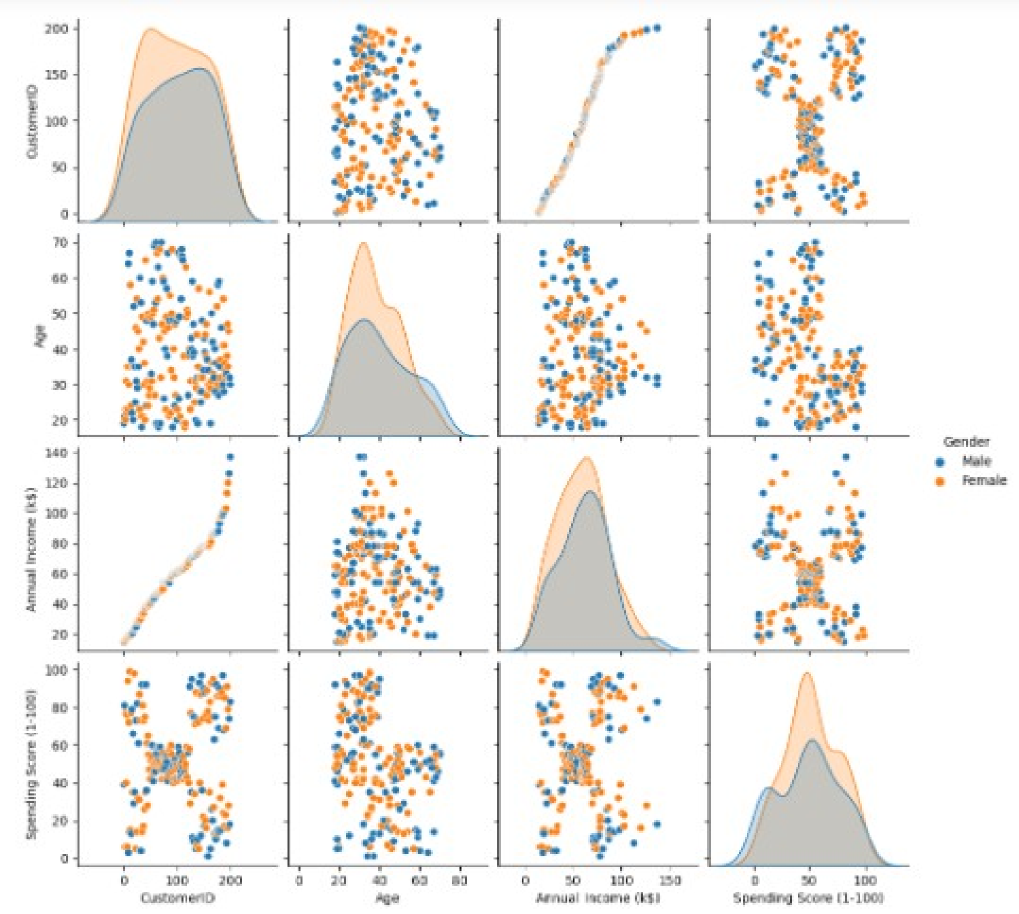
```
sns.scatterplot(data=mk,x='Annual Income (k$)', y='Spending 
```

Out[22]:

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

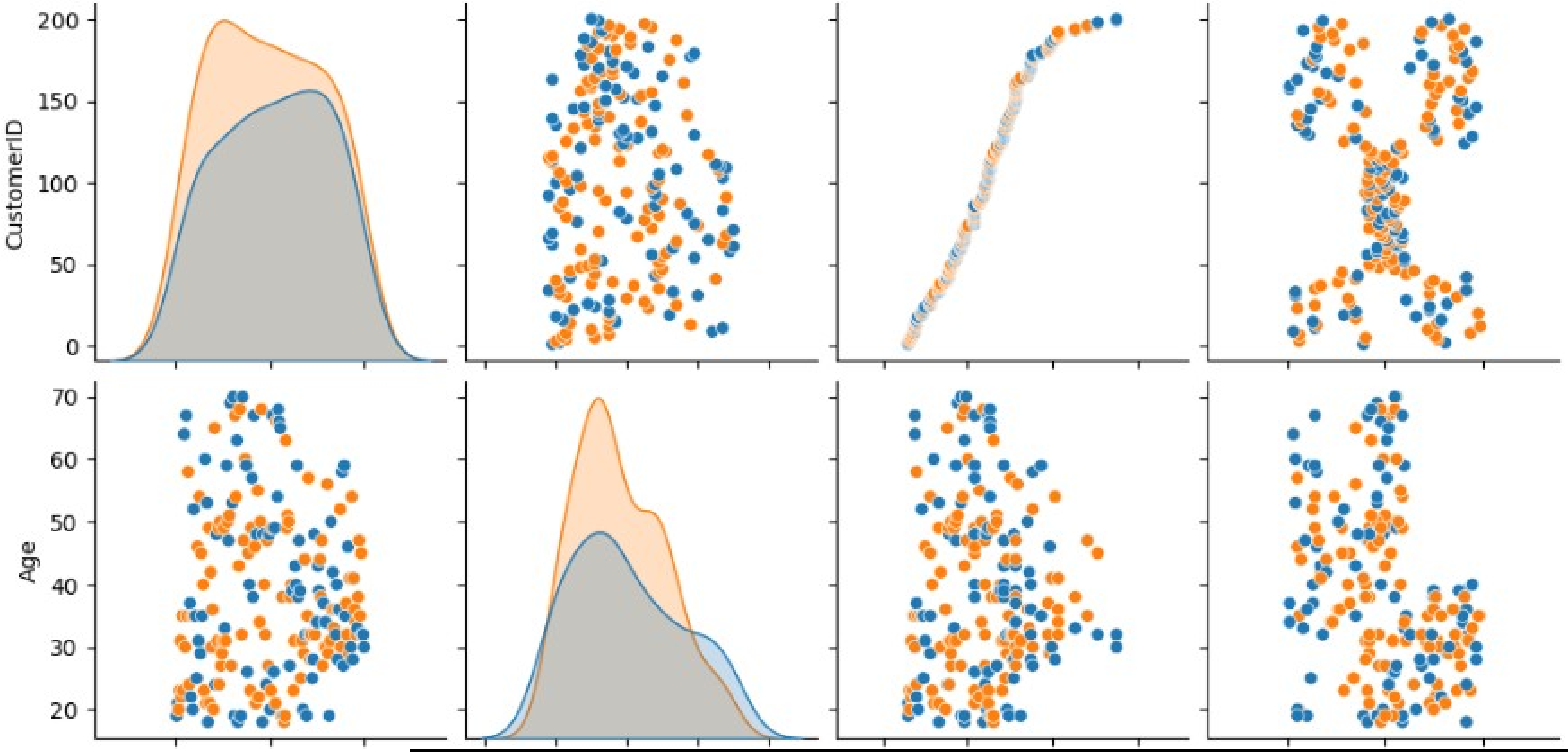


Bivariate analysis works with two variables
Using a scatter plot using seaborn analysis
In this scatter plot graphed by seaborn we set our X=Annual Income and Y=Spending Score and we can see some clusters between this two variables this is called bicluster variate we can see about 5-6 clusters.

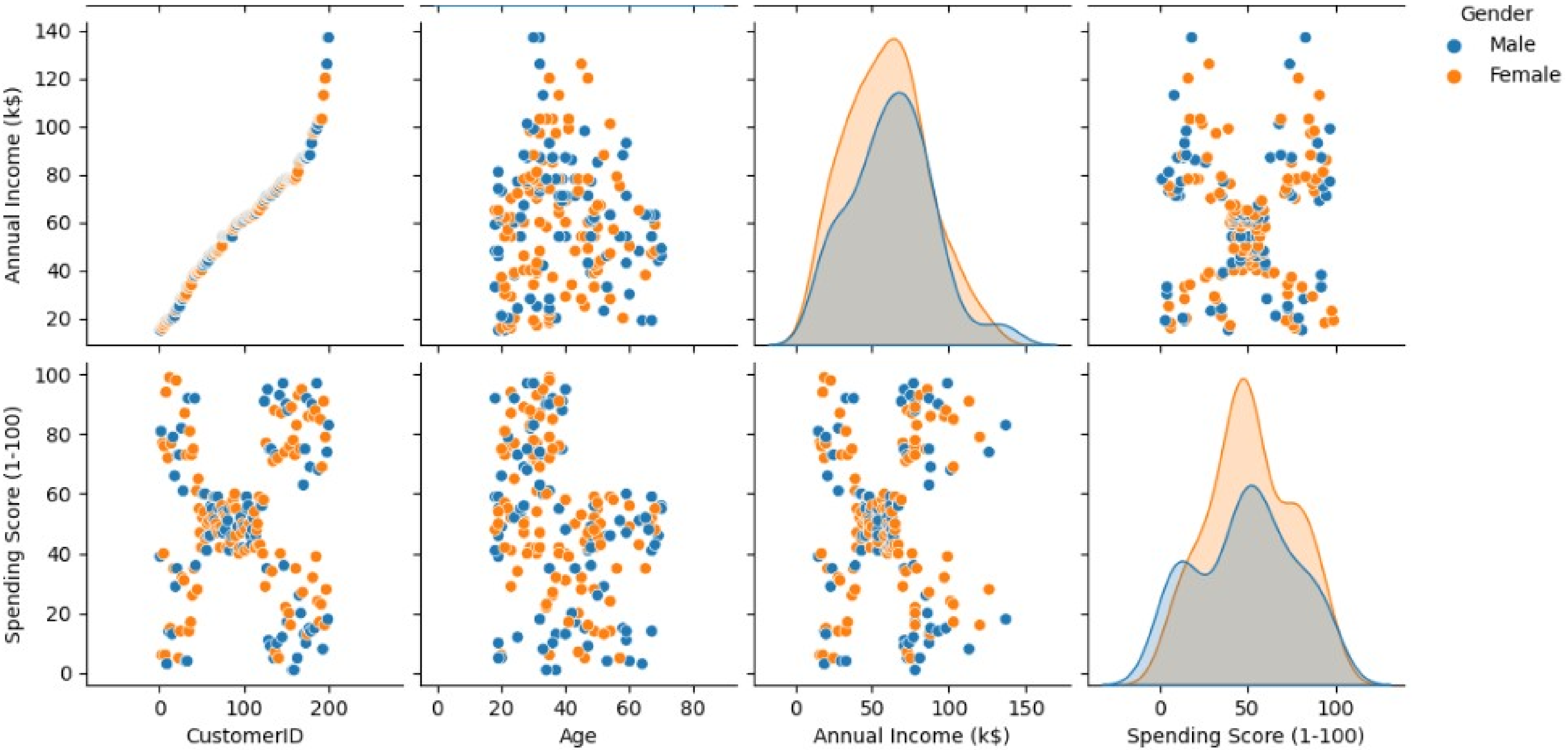


We would use hue to show the differences in relationship and clusters with gender

Bi Variate



Bi Variate



Gender
Male
Female
Annual Income (k$)
Spending Score (1-100)
CustomerID
Age
Annual Income (k$)
Spending Score (1-100)

Bi Variate

In [26]:

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

Out[26]:

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

We can see the mean values for our data using groupby to see mean values by gender.
We can see the age is similar, the annual income for males is higher as well, the spending score for males is also higher.

Bi Variate

In [27]:

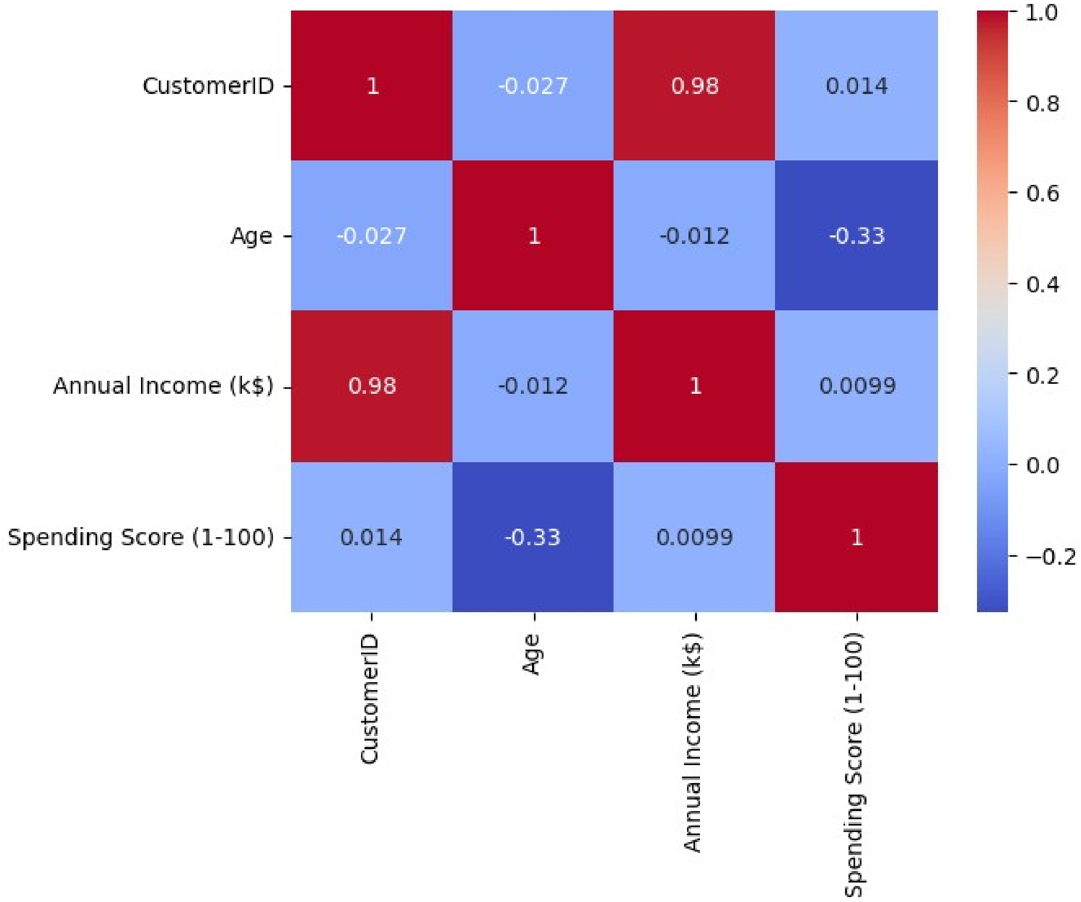
```
mk.corr()
```

Out[27]:

	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

We can look for the correlation for these two using correlation function(corr).
We can see with age annual income has a negative correlation, meaning as it increases it goes down.
Also spending score as well has a negative correlation with age.

Bi Variate



We can also use seaborn to create a heat map, and for parameters we can use annotations as a parameter, and also use c map which is the color mapping using cool and warm

Progress – Bi Variate Analysis

Analysis

- Used Bivariate analysis to explore our data and turned it into information we understand.
- Using a scatter plot using seaborn analysis
  In this scatter plot graphed by seaborn we set our X=Annual Income and Y=Spending Score and we can see some clusters between this two variables this is called bicluster variate we can see about 5-6 clusters.

- We used hue to show the differences in relationship and clusters with gender

Progress – Bi Variate Analysis

Analysis

- We got the mean values for our data using groupby to see mean values by gender.
- The age is similar, the annual income for male is higher as well, the spending score for male is also higher.
- We analyzed the mean values for our data using groupby to see mean values by gender
- We saw that the age is similar, the annual income for males is higher as well, the spending score for males is also higher.

Progress – Bi Variate Analysis

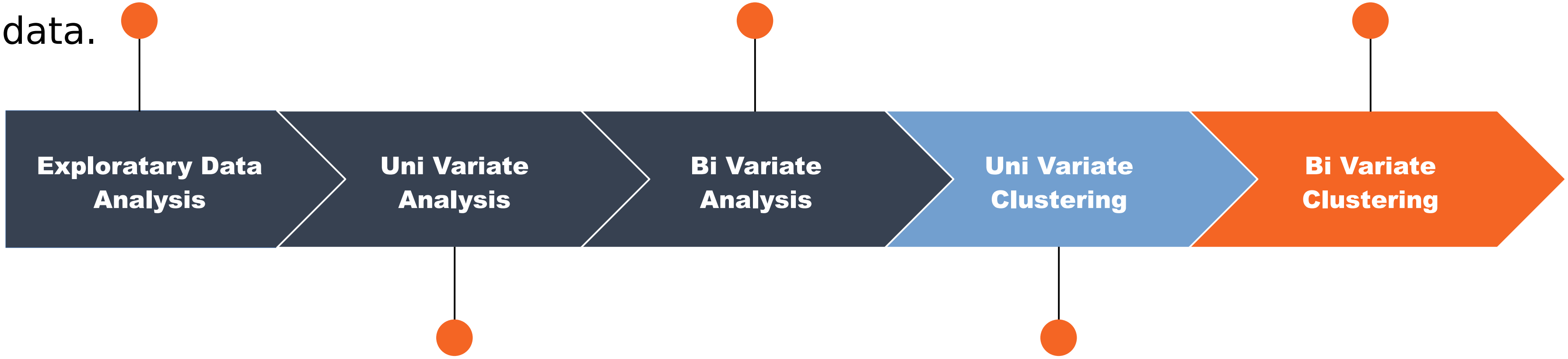
Analysis

- We looked for the correlation for these two using correlation function(corr).
  We saw that with age annual income has a negative correlation, meaning as it increases it goes down.
- Also spending score as well has a negative correlation with age.

Exploratary Data Analysis: Uncover surface level information about our data.

Bi Variate Analysis: Hue scatter plot.

Bi Variate Clustering:



Uni Variate Analysis: Histograms, KDE hue plot, box plot.

Uni Variate Clustering Analysis: mean values of age and spending score of the cluster using group by

Uni Variate Clustering Analysis

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

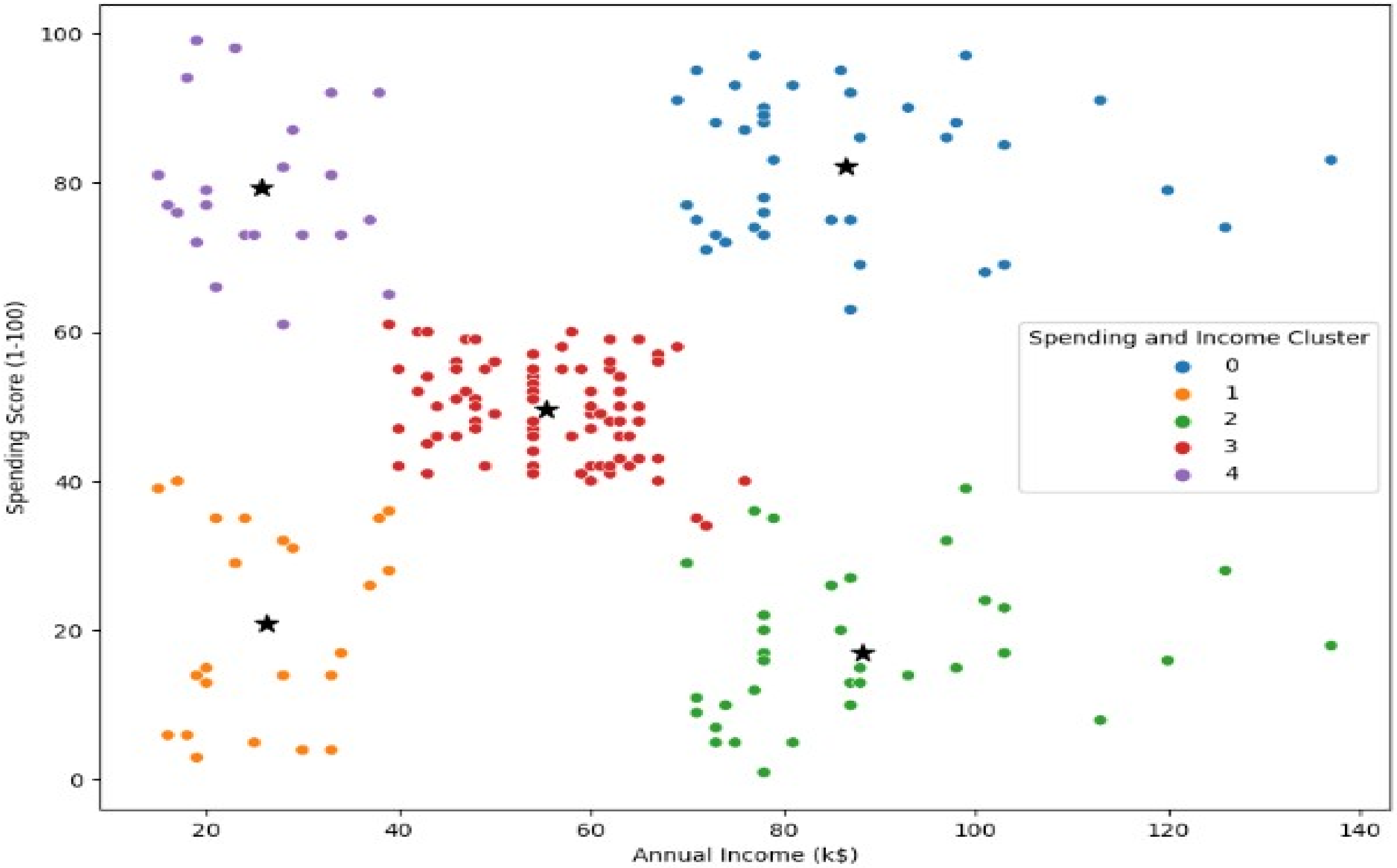
Out[27]:

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

The spending score is second lowest for the first cluster(0) and its annual income as well

Our second cluster(1) has the lowest for age but the highest annual income and spending score. Our third cluster(2) has the second highest for age and for annual income but the lowest for spending score.

Bi Variate Clustering Analysis



Bi Variate Clustering Analysis

[34]:
```
rosstab(mk['Spending and Income Cluster'],mk['Gender'],normalize='index')
```

t[34]:

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

We can see that cluster 0 which is the blue cluster has 53% female, cluster 1 the orange cluster with 59% is also dominated by females. The males in cluster 2 which is the green cluster with 54% are dominating. And in cluster 3 the red cluster is dominated by females with 59%. Finally the purple cluster which is 4 is al dominated by females 59%. From this analysis we can see that the males only dominated cluster 2 while the females dominated 4 clusters.

So our ideal cluster which would be the high spending score and annual income would be cluster 0 would be our target cluster because that cluster would bring the most money.

Bi Variate Clustering Analysis

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

]:

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

In cluster 0 we have high annual income and spending score with a low age but not the lowest age the lowest is cluster 4, they have a low annual income but a high spending score we can assume they are coming in for a big ticket item so we can build a campaign around that, using the customer id we can see the items purchased.
We can say from the data cluster 0 which is 53% female has a high spending score a high annual income and a low age of 32% is our ideal cluster to run campaigns on.

Final Analysis

1. Our target group would be cluster 0 which has ahigh spending score and a high income
2. cluster 0 has 53% female shoppers, we could create marketing campaigns around them using popular items bought by them
3. We should also target cluster 4, they have a low annual income but a high spending score we can assume they are coming in for a big ticket item so we can build a campaign around that, using the customer id we can see the items purchased.