DATE - 16\10\2023

TEAM ID - 3884

PROJECT TITLE - Age Based Customer Segmentation using Data Science

Importing Dependencies

In [64]:

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

Loading Dataset

In [65]:

```
import pandas as pd
df = pd.read_csv("C:\\Users\\sowen\\OneDrive\\Documents\\phase 3 customer s
df
```

Out[65]:

	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

Data Exploration

In [66]:
```
print(df.head())
```

```
   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
```

In [67]:
```
print(df.tail(10))
```

```
     CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)
190         191  Female   34                 103                      23
191         192  Female   32                 103                      69
192         193    Male   33                 113                       8
193         194  Female   38                 113                      91
194         195  Female   47                 120                      16
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
```

In [68]:
```
dataset
```

Out[68]:

	CustomerID	Genre	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

In [69]:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   CustomerID              200 non-null    int64 
 1   Genre                   200 non-null    object
 2   Age                     200 non-null    int64 
 3   Annual Income (k$)      200 non-null    int64 
 4   Spending Score (1-100)  200 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

In [70]:

```
dataset.describe()
```

Out[70]:

	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 [71]:

```
dataset.columns
```

Out[71]:

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

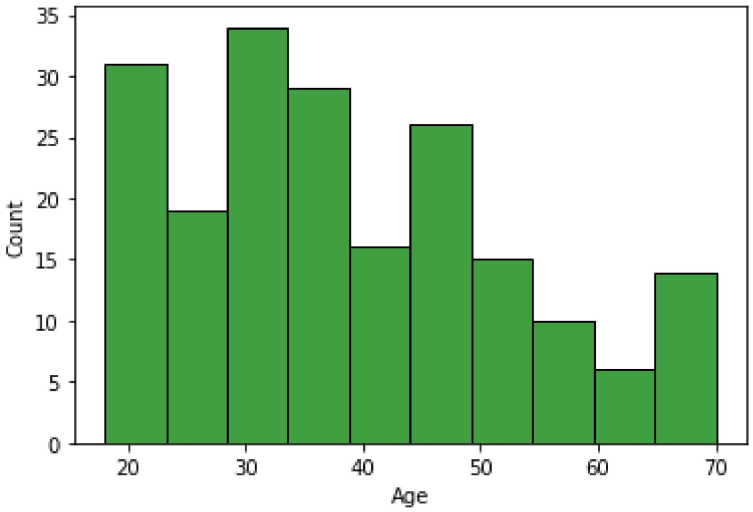
Pre-Processing and visualisation of data

In [72]:

```
sns.histplot(dataset, x='Age', bins=10, color='g')
```

Out[72]:

```
<AxesSubplot:xlabel='Age', ylabel='Count'>
```



Check for missing values

In [73]:

```
print(df.isnull().sum())
```

```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

In [74]:

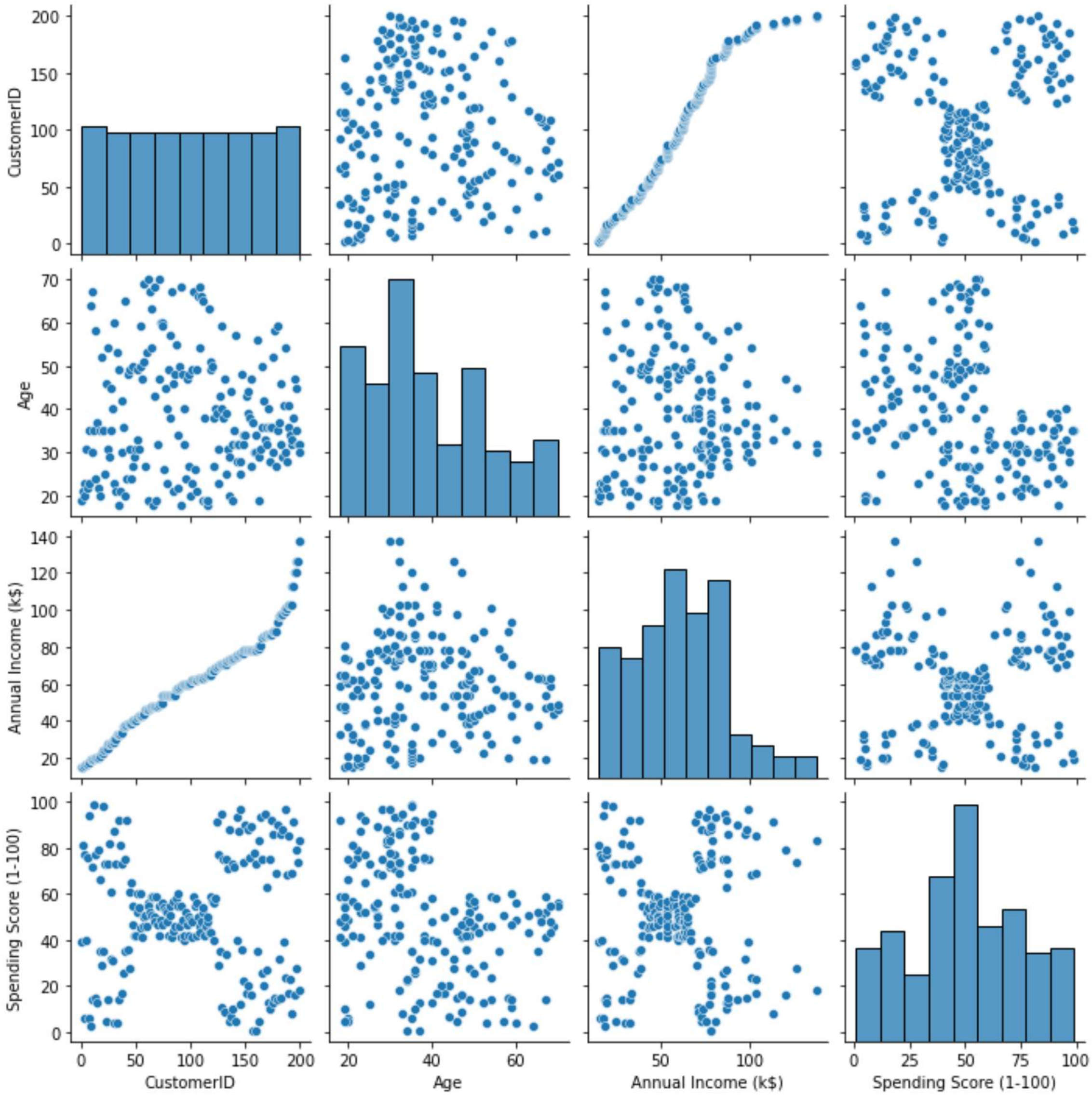
```
print(df.dropna().sum())
```

```
CustomerID                                                            2010
0
Gender                    MaleMaleFemaleFemaleFemaleFemaleFemaleFemaleM
a...
Age                                                                    777
0
Annual Income (k$)                                                    1211
2
Spending Score (1-100)                                                1004
0
dtype: object
```

```
In [75]: plt.figure(figsize=(12,8))
         sns.pairplot(dataset)
```

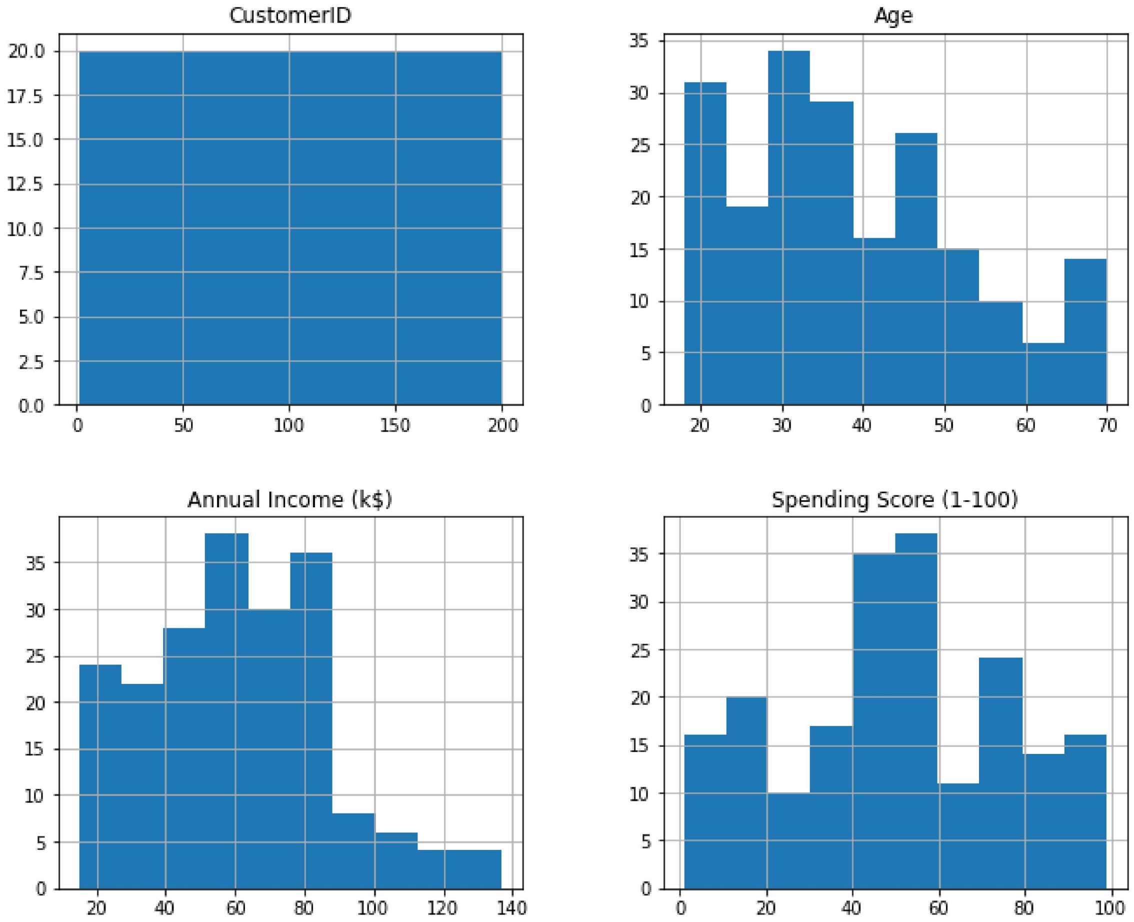
```
Out[75]: <seaborn.axisgrid.PairGrid at 0x2d0706a7f40>

         <Figure size 864x576 with 0 Axes>
```



```
In [76]: dataset.hist(figsize=(11,9))
```

```
Out[76]: array([[<AxesSubplot:title={'center':'CustomerID'}>,
                 <AxesSubplot:title={'center':'Age'}>],
                [<AxesSubplot:title={'center':'Annual Income (k$)'}>,
                 <AxesSubplot:title={'center':'Spending Score (1-100)'}>]],
               dtype=object)
```



Visualising correlation

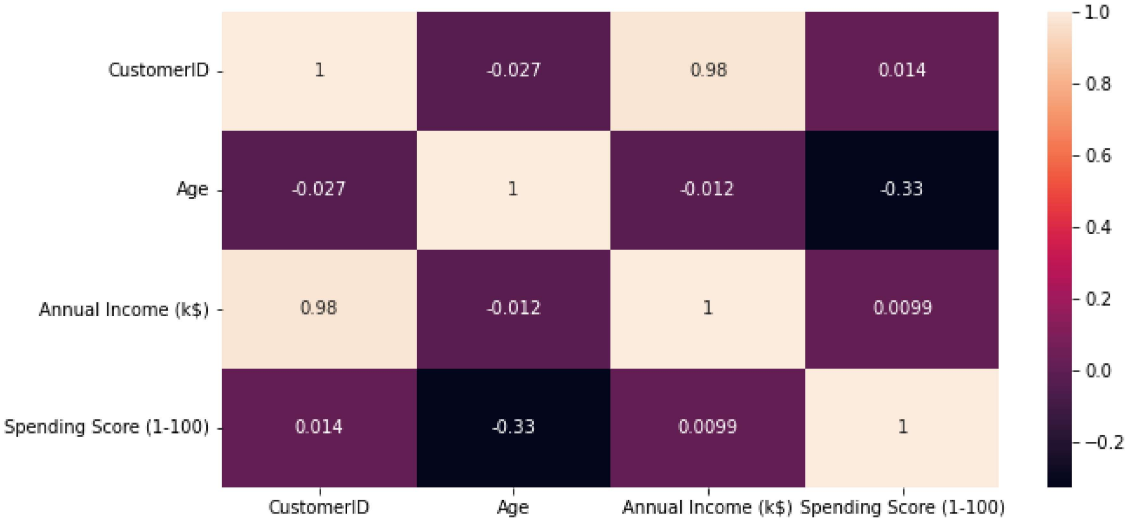
```
In [77]: dataset.corr()
```

Out[77]:

	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 [78]: plt.figure(figsize=(10,5))
         sns.heatmap(dataset.corr(), annot=True)
```

```
Out[78]: <AxesSubplot:>
```


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