SYNTHETIC CUSTOMER PROFILING DATA

This report analyzes the customer profiling scenario of retail company for e.g. what age os customers has highest score , is there any relationship betwen customer age and their purchases ? etc.

DATASET OVERVIEW

- **ROWS**: 200
- **Columns**: 5
- **Columns Names**: ID, Age, Income in Rs., Score, Purchases
  - `Id` =Unique identifier for each customers
  - `Age` = age of customers
  - `Income in Rs.` =estimated annual income of each customers
  - `Score` =a score from 0 to 100 indicating customer's layalty towards company
  - `Purchases` =number of purchases by each customers in last 6 months or 1 year

DATA CLEANING SUMARRY

- Filled 2 empty cells from column- 'Income', 'Score'
- Correct an outliar from 'Score' column,code used-
  - `print(np.where((df['Score']>100) | (df['Score']<0)))`
  - `df.loc[75,'Score']=100`
  - `df['Score'].fillna(df['Score'].mean().round(2),inplace=True)`
- Converted negative values of 'Age' into positive value
- Fixed data types to int

EXPLORATORY DATA ANALYSIS

- AVG age of customers are 39,while the minimum age is 18 and maximum age is 59
- AVG annual income of customers are Rs. 52691.97, where around half of the total customers have income around Rs.42266.75
- AVG scores of customers are 50 , with a max score of 100 and minimum of 1
- AVG purchases made in 6 months or 1 year by each customer is around 5 units

In [5]:
```
import pandas as pd, numpy as np, matplotlib.pyplot as plt ,seaborn as sns
df=pd.read_csv('cleaned_file.csv')
df.drop('Unnamed: 0',axis=1,inplace=True)
print(df.describe().round(2))
```

```
            ID     Age  Income in Rs.   Score  Purchases
count   200.00  200.00         200.00  200.00     200.00
mean    100.50   38.72       52691.97   50.44       4.99
std      57.88   12.57       14729.71   31.02       2.20
min       1.00   18.00       20075.00    1.00       1.00
25%      50.75   28.00       42266.75   20.75       3.00
50%     100.50   40.00       52057.50   51.50       5.00
75%     150.25   49.25       63876.25   79.00       7.00
max     200.00   59.00       91016.00  100.00      11.00
```

- There is no correleation between any of the variables

In [6]:
```
print(df.corr().round())
```

```
                ID  Age  Income in Rs.  Score  Purchases
ID             1.0  0.0           -0.0   -0.0        0.0
Age            0.0  1.0           -0.0    0.0        0.0
Income in Rs. -0.0 -0.0            1.0   -0.0        0.0
Score         -0.0  0.0           -0.0    1.0        0.0
Purchases      0.0  0.0            0.0    0.0        1.0
```

- Group wise stats
  - grouping by 'Age' and calcluated mean and maximum value of variables 'Income in Rs.','Score'
  - grouping by 'Score' and calculated minimum and maximum value of variables 'Purchases'

In [7]:
```
group_1=df.groupby('Age')[['Income in Rs.','Score']].agg(['mean','max']).round(2)
print(group_1)
group_2=df.groupby('Score')['Purchases'].agg(['max','min'])
print("\n\n",group_2)
```

Age	Income in Rs.		Score	
	mean	max	mean	max
18	55763.50	77651	75.00	99
19	46458.33	68140	46.89	89
20	49643.60	58045	54.60	98
21	56901.00	77655	23.60	55
22	47027.50	53484	65.50	86
23	54406.75	65978	53.50	87
24	45579.00	68256	48.60	98
25	59695.14	71560	35.86	72
26	51428.20	70730	25.20	67
27	50629.00	50629	69.00	69
28	61609.20	75599	54.20	88
29	40138.67	57896	47.00	66
30	59240.33	68996	49.00	70
31	50369.17	65683	52.00	95
32	54960.88	71078	54.38	97
33	64255.50	67860	49.50	62
34	49467.75	73292	74.25	96
35	44205.33	51387	41.33	96
36	52703.50	62809	28.50	54
37	62597.50	63646	54.50	58
38	48047.00	58341	41.40	90
39	59531.50	73146	33.00	86
40	51307.40	63847	61.60	95
41	53089.44	71788	46.33	91
42	65862.00	88087	55.20	86
43	43446.83	76444	50.17	100
44	59186.00	81669	54.00	94
45	59333.88	85219	60.62	96
46	50383.75	70760	54.00	79
47	43392.33	46963	72.00	82
48	61105.00	61105	58.00	58
49	43632.38	78684	46.38	92
50	58729.00	69208	58.50	98
51	60008.00	64657	79.00	96
52	64458.00	85835	61.50	99
53	40363.50	47940	30.50	62
54	52790.29	67429	43.86	87
55	36717.00	36717	70.00	70
56	46036.40	78080	41.60	94
57	52738.25	63180	60.00	96
58	43673.25	58276	46.75	58
59	60340.83	91016	53.67	99

Score	max	min
1	6	1
2	3	2
3	6	6
4	10	1
5	5	4
...	...	...
96	8	1
97	5	5
98	8	5
99	5	1
100	10	10

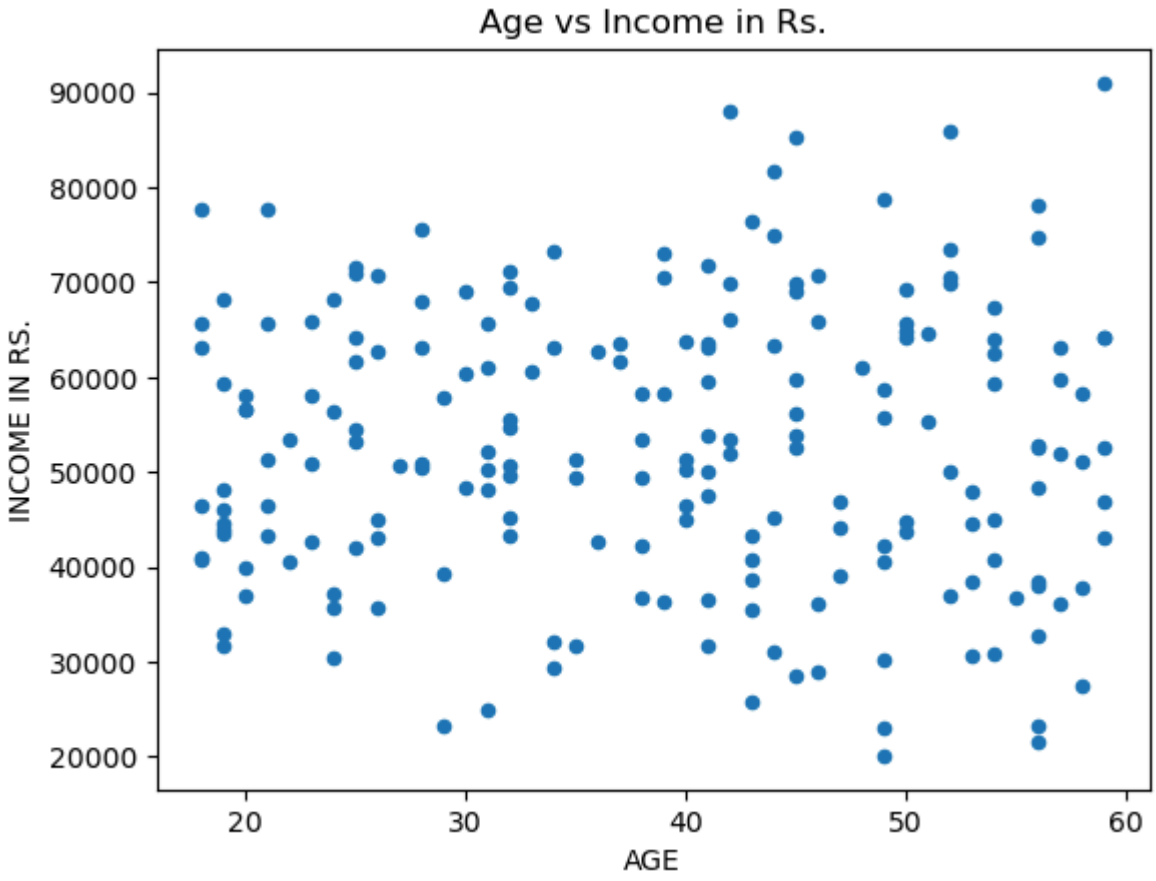
[90 rows x 2 columns]

DATA VISUALIZATION

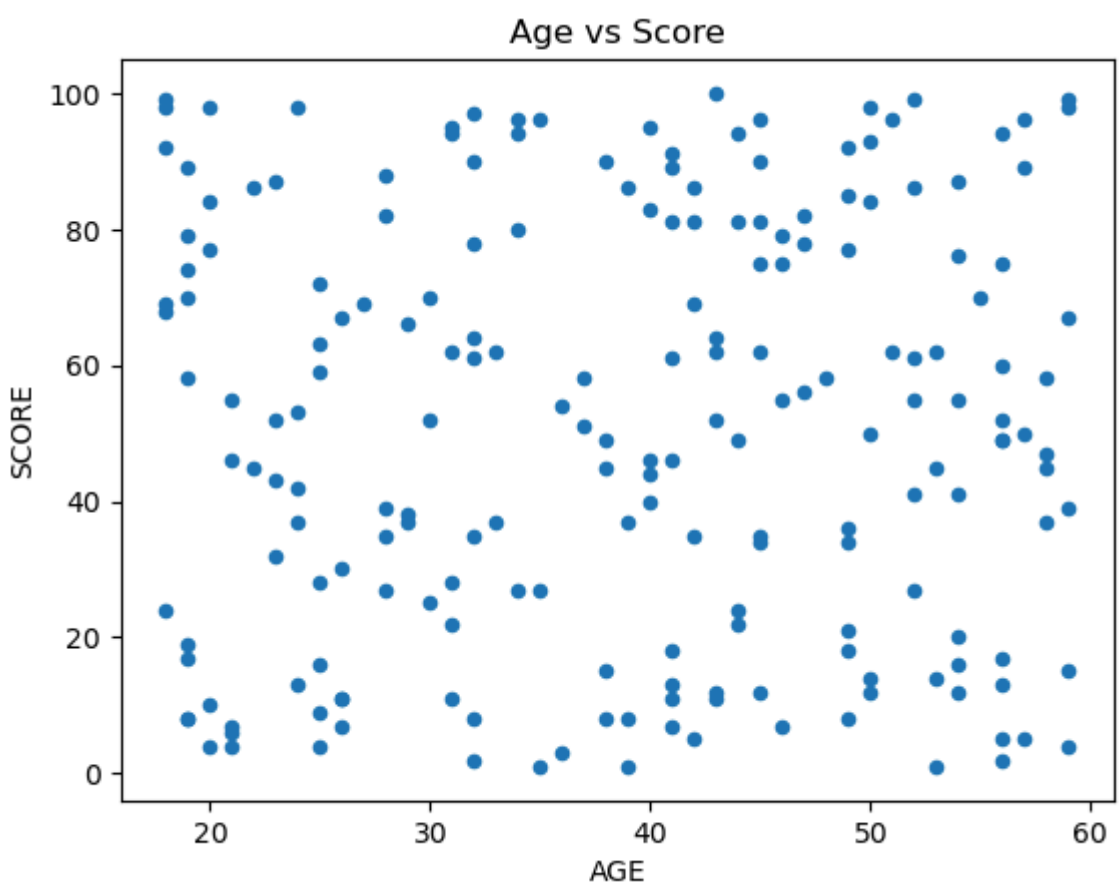
- *Age vs INCOME* (scatter)
- *AGE vs Score* (scatter)
- *Score vs Purchase* (scatter)
- *Age Distribution* (frequency distribution)
- *Age vs purchase* (scatter)

In [8]:
```
print('Age vs Income in Rs.')
x=df.sort_values(by='Age')                                    # sorting by age in a variabe
x.plot(kind='scatter',x='Age',y='Income in Rs.' , xlabel= 'AGE',ylabel='INCOME IN RS.')
plt.title('Age vs Income in Rs.')
plt.show()


print('Age vs Score')
x.plot(kind='scatter',x='Age',y='Score' , xlabel= 'AGE',ylabel='SCORE')
plt.title('Age vs Score')

plt.show()

print('Score vs Purchases')
y=df.sort_values(by='Score')                                  # sorting by Score in a varia
y.plot(kind='scatter',x='Purchases',y='Score' , xlabel= 'PURCHASES',ylabel='SCORE')
plt.title('Score vs Purchases')
plt.show()

print('Age vs Purchases')
x.plot(kind='scatter',x='Age',y='Purchases' , xlabel= 'AGE',ylabel='Purchases')
plt.title('Age vs Purchases')
plt.show()

print('Age Distribution')
df['Age'].plot(kind='kde',xlabel= 'AGE',ylabel='FREQUENCY')
plt.title('Age Distribution')
plt.show()
```

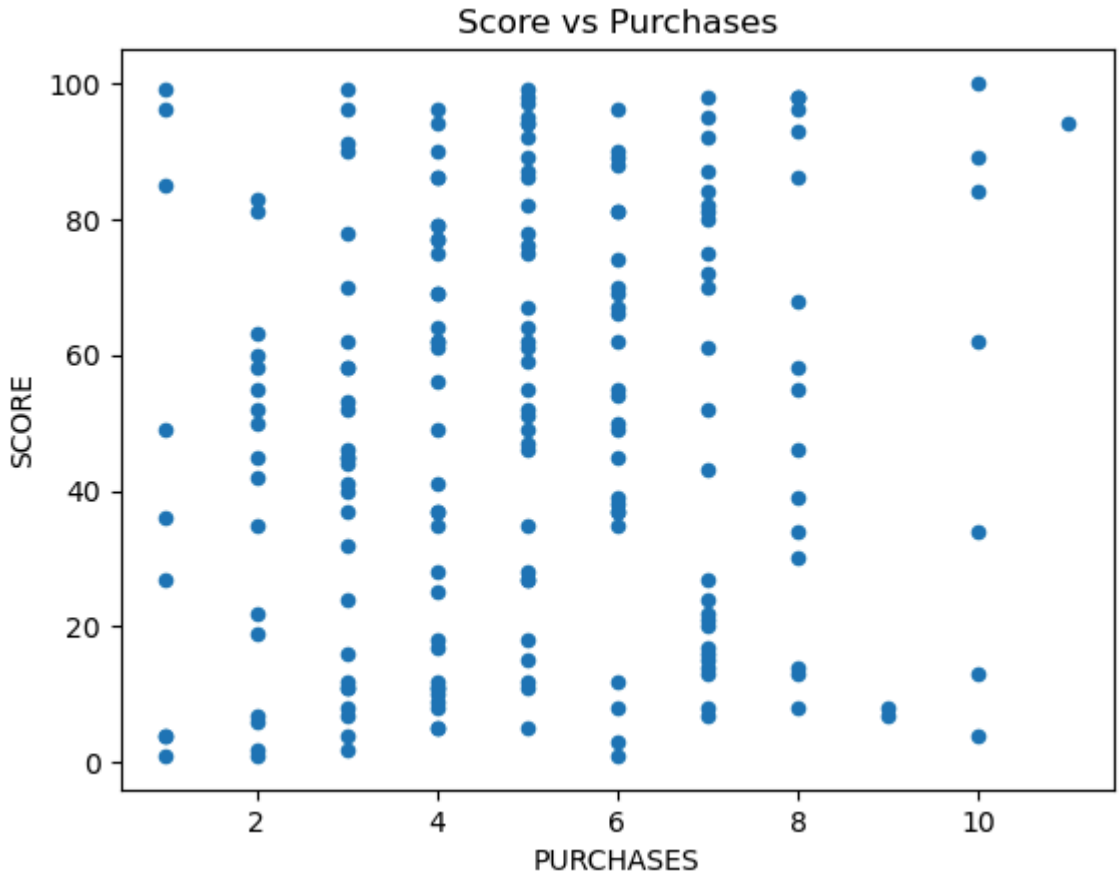
```
Age vs Income in Rs.
```



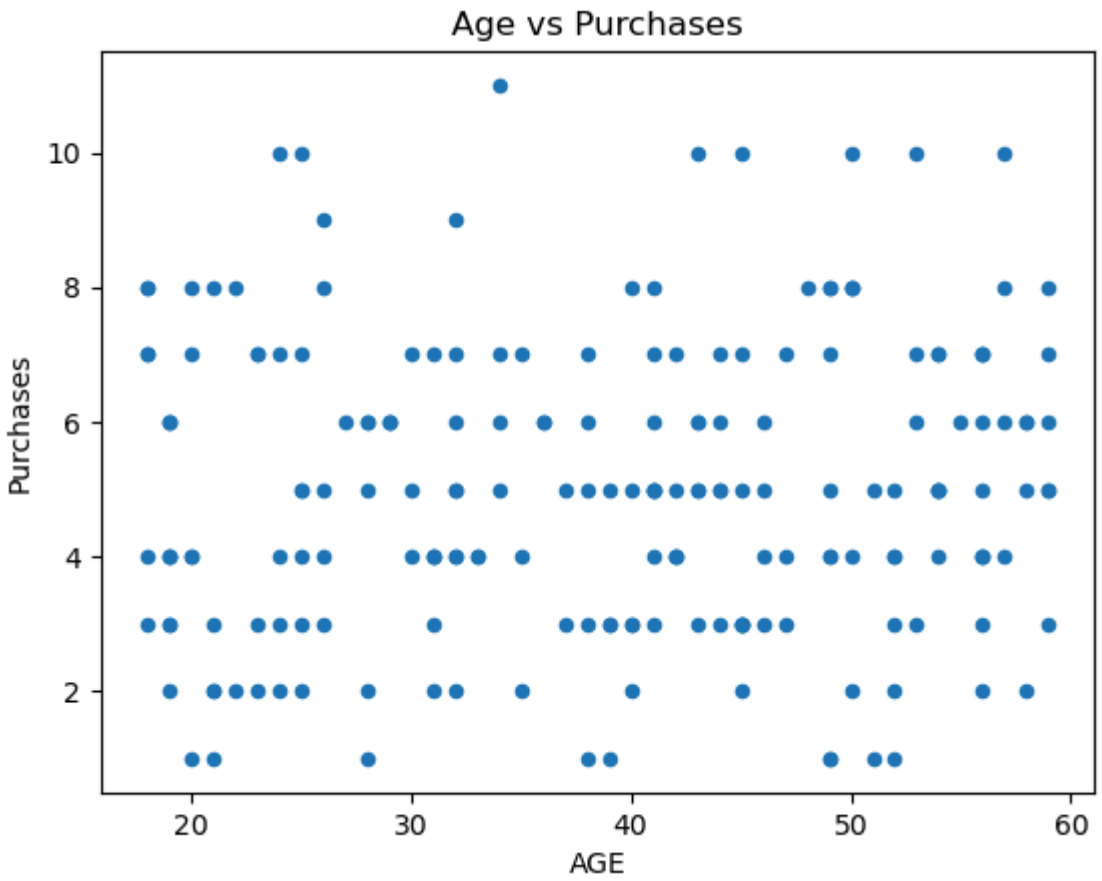
Age vs Score



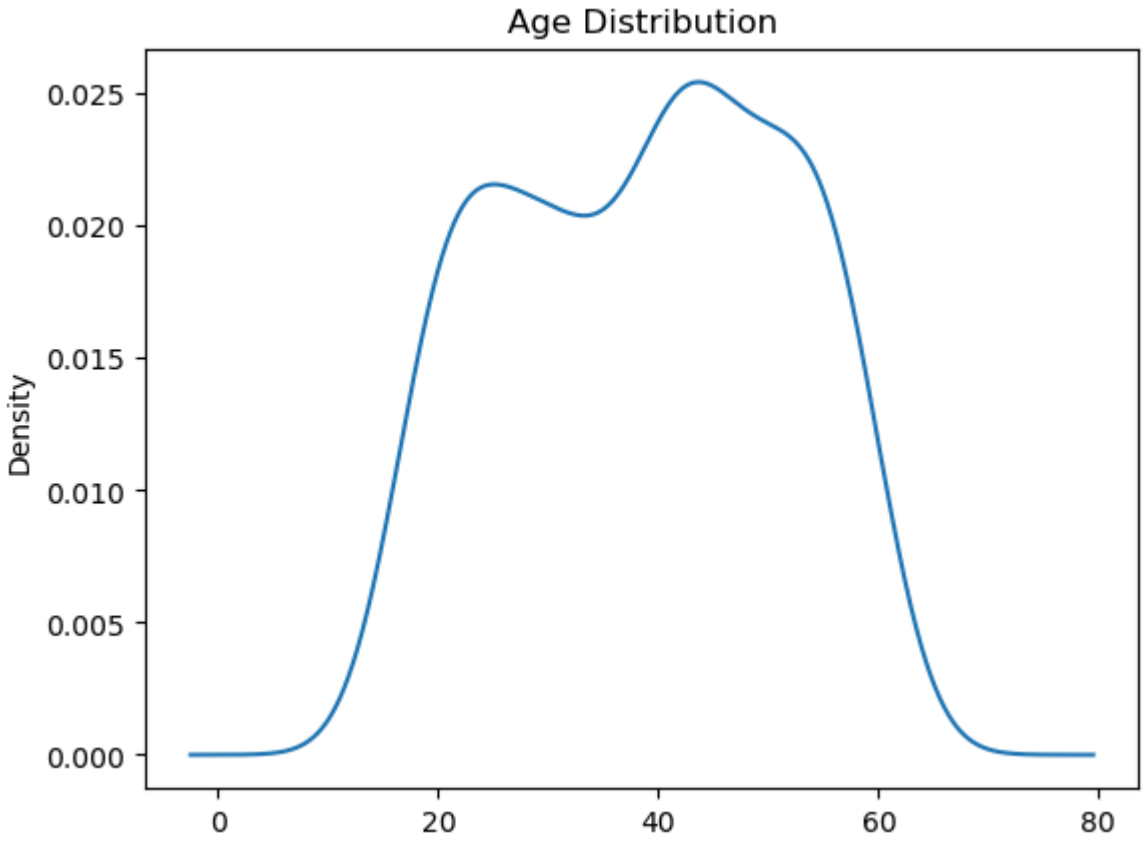
Score vs Purchases



Age vs Purchases



Age Distribution



KEY INSIGHTS

- The company has over half of his customers of age around 40
- The wealthiest customer has annual income of Rs. 91016 , with an avg income of their customers Rs. 52691.97
- On an average customers scored 50 in company's customer loyalty list
- In a year or in 6 months , the maximum sale has gone by a single customer is 11 , while avg sales by [illegible] a minimum of 1 unit sold to a customer

- From the Scatter Graphs and Correlation formula we found tha there is no association between any variables