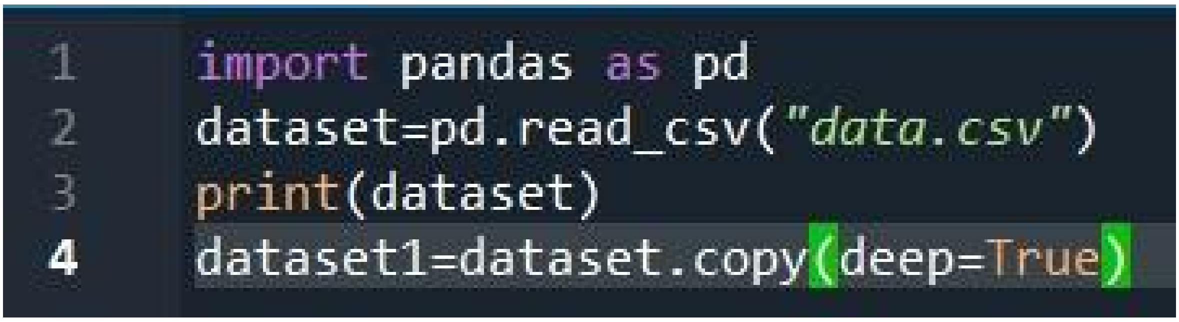
Pooja Agarwal
1905330

Lab 2

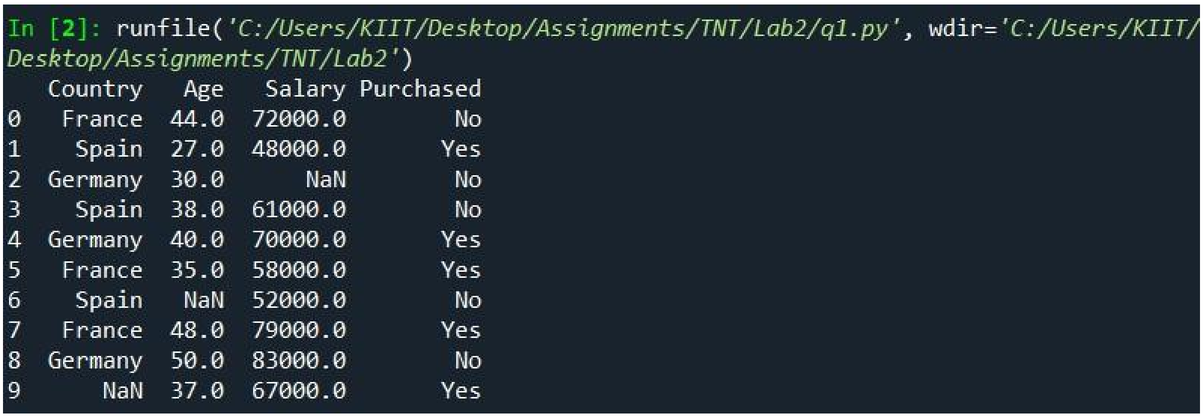
Q1: Create a copy of the ‘data.csv’ and name the dataframe as dataset1

CODE:



```
import pandas as pd
dataset=pd.read_csv("data.csv")
print(dataset)
dataset1=dataset.copy(deep=True)
```

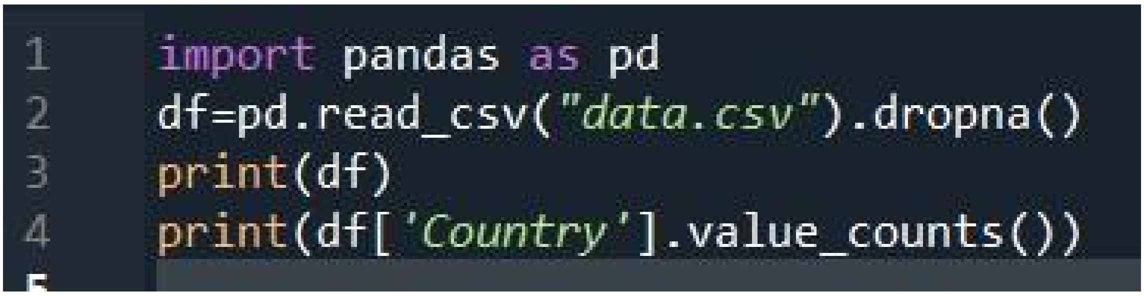
OUTPUT:



```
In [2]: runfile('C:/Users/KIIT/Desktop/Assignments/TNT/Lab2/q1.py', wdir='C:/Users/KIIT/
Desktop/Assignments/TNT/Lab2')
   Country   Age   Salary Purchased
0   France  44.0  72000.0        No
1    Spain  27.0  48000.0       Yes
2  Germany  30.0      NaN        No
3    Spain  38.0  61000.0        No
4  Germany  40.0  70000.0       Yes
5   France  35.0  58000.0       Yes
6    Spain   NaN  52000.0        No
7   France  48.0  79000.0       Yes
8  Germany  50.0  83000.0        No
9      NaN  37.0  67000.0       Yes
```

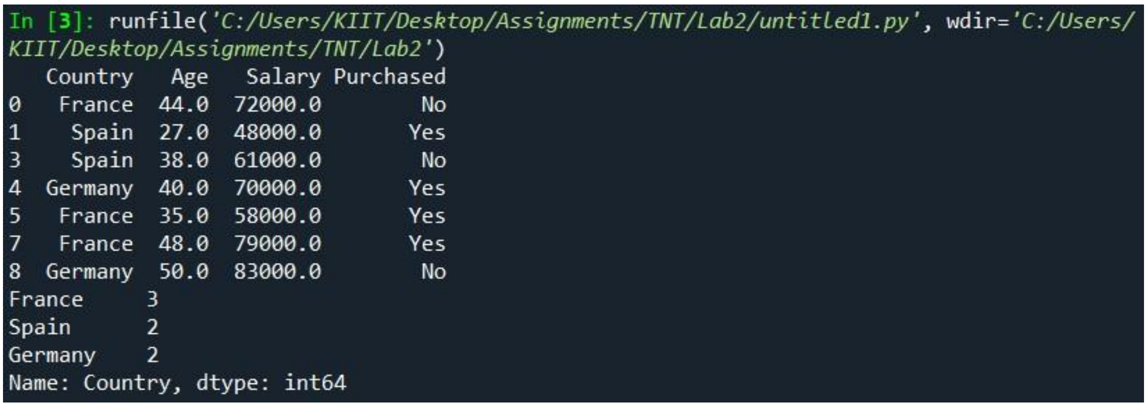
Q2: To display the count of each value in the county column

CODE:



```
import pandas as pd
df=pd.read_csv("data.csv").dropna()
print(df)
print(df['Country'].value_counts())
```

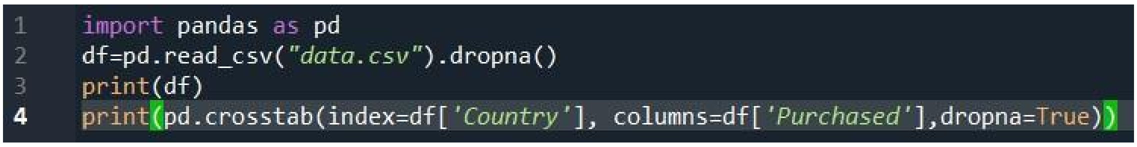
OUTPUT:



```
In [3]: runfile('C:/Users/KIIT/Desktop/Assignments/TNT/Lab2/untitled1.py', wdir='C:/Users/
KIIT/Desktop/Assignments/TNT/Lab2')
   Country   Age   Salary Purchased
0   France  44.0  72000.0        No
1    Spain  27.0  48000.0       Yes
3    Spain  38.0  61000.0        No
4  Germany  40.0  70000.0       Yes
5   France  35.0  58000.0       Yes
7   France  48.0  79000.0       Yes
8  Germany  50.0  83000.0        No
France     3
Spain      2
Germany    2
Name: Country, dtype: int64
```

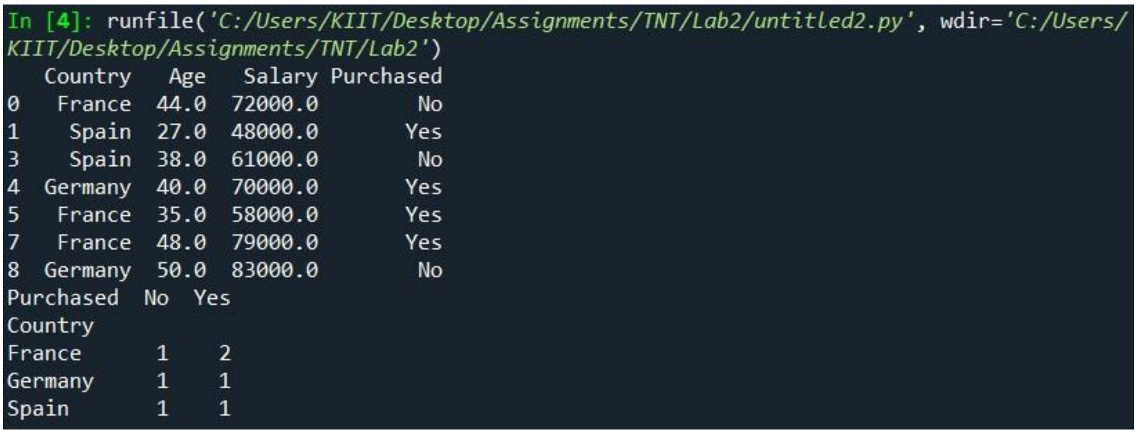
Q3: To display how many individuals from each country are buying the product and how many aren't.

CODE:



```
import pandas as pd
df=pd.read_csv("data.csv").dropna()
print(df)
print(pd.crosstab(index=df['Country'], columns=df['Purchased'],dropna=True))
```

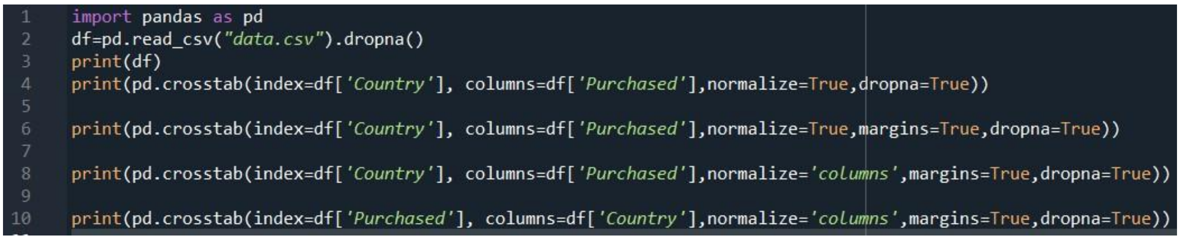
OUTPUT:



```
In [4]: runfile('C:/Users/KIIT/Desktop/Assignments/TNT/Lab2/untitled2.py', wdir='C:/Users/
KIIT/Desktop/Assignments/TNT/Lab2')
   Country   Age   Salary Purchased
0   France  44.0  72000.0        No
1    Spain  27.0  48000.0       Yes
3    Spain  38.0  61000.0        No
4  Germany  40.0  70000.0       Yes
5   France  35.0  58000.0       Yes
7   France  48.0  79000.0       Yes
8  Germany  50.0  83000.0        No
Purchased  No  Yes
Country
France      1    2
Germany     1    1
Spain       1    1
```

Q4: Show all probabilities of occurance:
i) Joint
ii) Marginal
iii) Conditional:
A) Country is known, whether the individual will purchase the product or not
B) Product has been brought/not brought, find the probability the individual belongs to which country

CODE:



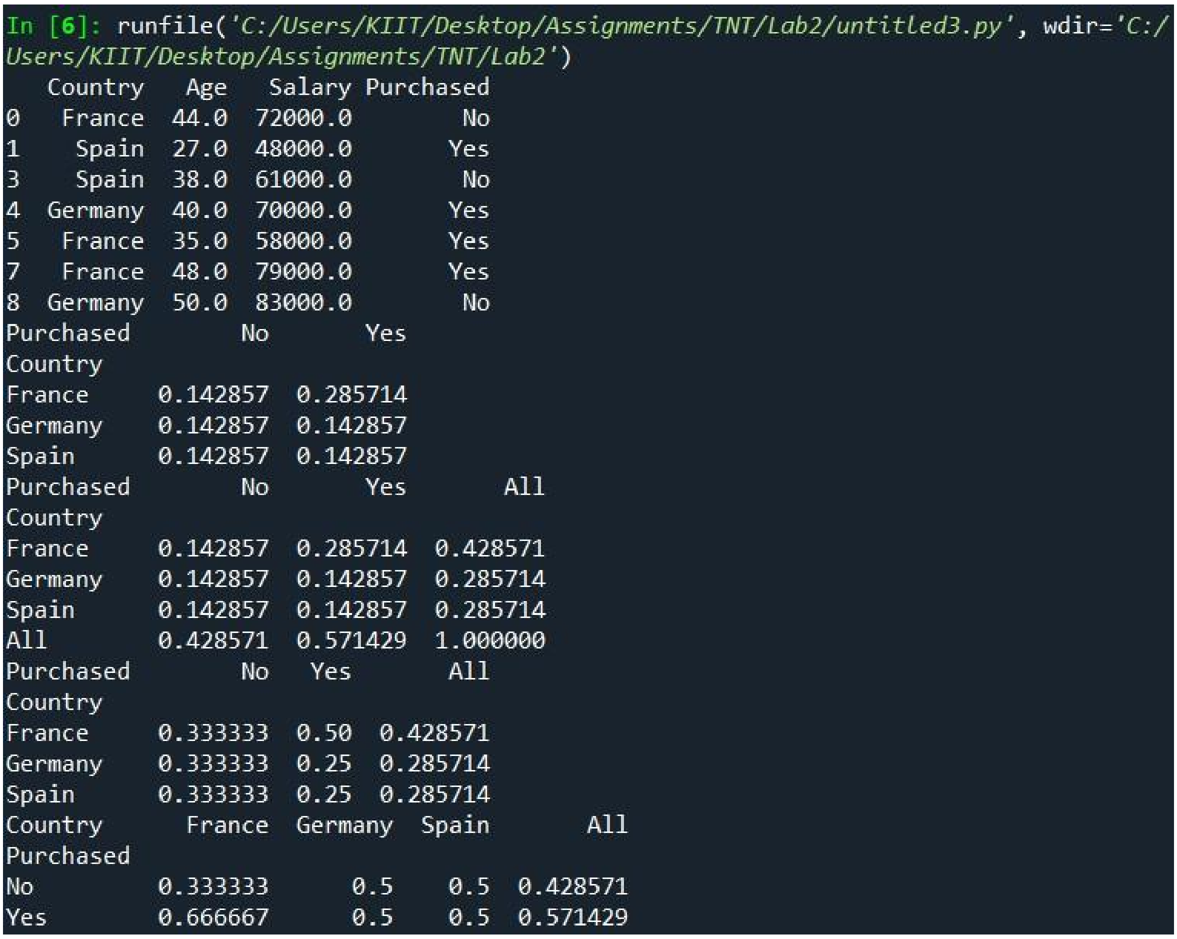
```
import pandas as pd
df=pd.read_csv("data.csv").dropna()
print(df)
print(pd.crosstab(index=df['Country'], columns=df['Purchased'],normalize=True,dropna=True))

print(pd.crosstab(index=df['Country'], columns=df['Purchased'],normalize=True,margins=True,dropna=True))

print(pd.crosstab(index=df['Country'], columns=df['Purchased'],normalize='columns',margins=True,dropna=True))

print(pd.crosstab(index=df['Purchased'], columns=df['Country'],normalize='columns',margins=True,dropna=True))
```

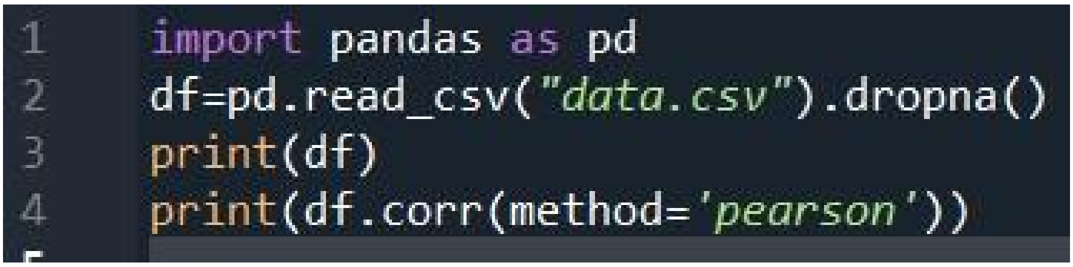
OUTPUT:



```
In [6]: runfile('C:/Users/KIIT/Desktop/Assignments/TNT/Lab2/untitled3.py', wdir='C:/
Users/KIIT/Desktop/Assignments/TNT/Lab2')
   Country   Age   Salary Purchased
0   France  44.0  72000.0        No
1    Spain  27.0  48000.0       Yes
3    Spain  38.0  61000.0        No
4  Germany  40.0  70000.0       Yes
5   France  35.0  58000.0       Yes
7   France  48.0  79000.0       Yes
8  Germany  50.0  83000.0        No
Purchased        No       Yes
Country
France     0.142857  0.285714
Germany    0.142857  0.142857
Spain      0.142857  0.142857
Purchased        No       Yes       All
Country
France     0.142857  0.285714  0.428571
Germany    0.142857  0.142857  0.285714
Spain      0.142857  0.142857  0.285714
All        0.428571  0.571429  1.000000
Purchased        No   Yes       All
Country
France     0.333333  0.50  0.428571
Germany    0.333333  0.25  0.285714
Spain      0.333333  0.25  0.285714
Country      France  Germany  Spain       All
Purchased
No         0.333333      0.5    0.5  0.428571
Yes        0.666667      0.5    0.5  0.571429
```

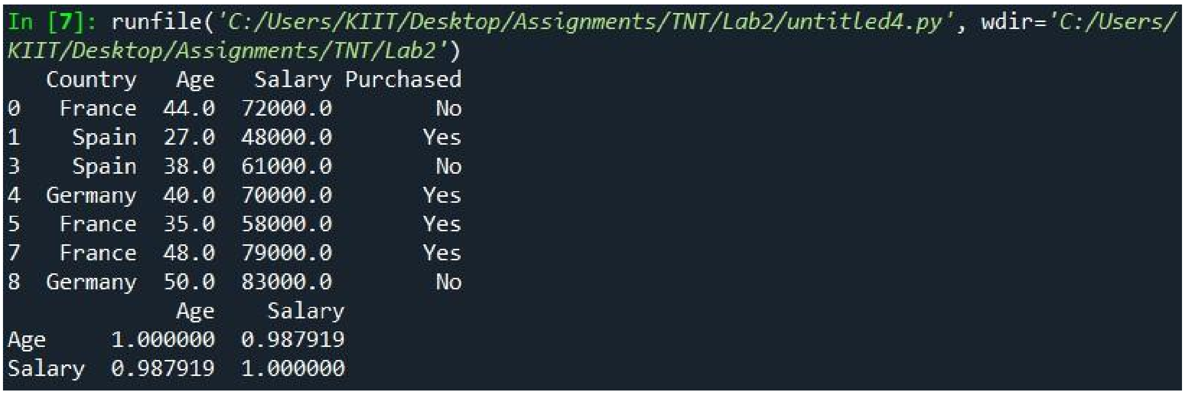
Q5: Find out whether there is a correlation between numerical data(variables) in the dataset.

CODE:



```
import pandas as pd
df=pd.read_csv("data.csv").dropna()
print(df)
print(df.corr(method='pearson'))
```

OUTPUT:

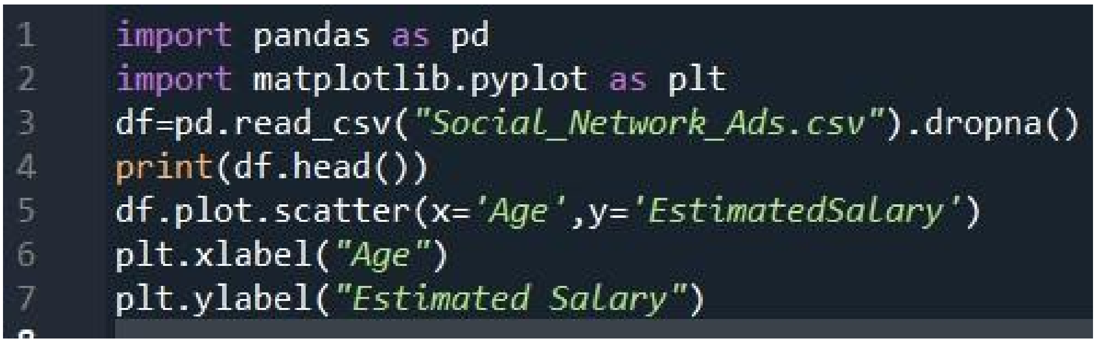


```
In [7]: runfile('C:/Users/KIIT/Desktop/Assignments/TNT/Lab2/untitled4.py', wdir='C:/Users/
KIIT/Desktop/Assignments/TNT/Lab2')
   Country   Age   Salary Purchased
0   France  44.0  72000.0        No
1    Spain  27.0  48000.0       Yes
3    Spain  38.0  61000.0        No
4  Germany  40.0  70000.0       Yes
5   France  35.0  58000.0       Yes
7   France  48.0  79000.0       Yes
8  Germany  50.0  83000.0        No
             Age    Salary
Age     1.000000  0.987919
Salary  0.987919  1.000000
```

Pooja Agarwal
1905330

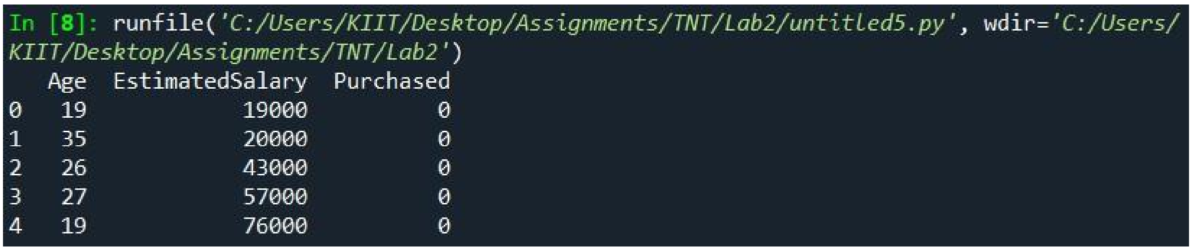
Q6: Use scatter plot and plot the data given in 'social_network_ad.csv'.Keep 'age' in x-axis and 'estimated salary' in y-axis.

CODE:



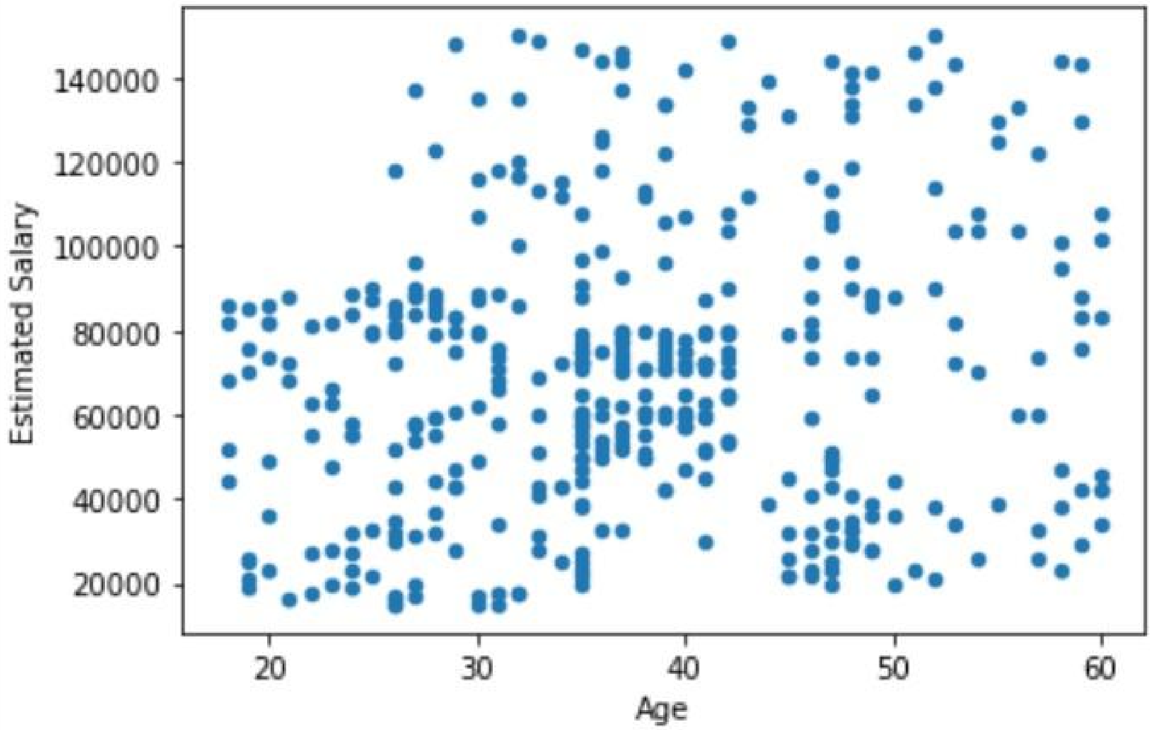
```
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv("Social_Network_Ads.csv").dropna()
print(df.head())
df.plot.scatter(x='Age',y='EstimatedSalary')
plt.xlabel("Age")
plt.ylabel("Estimated Salary")
```

OUTPUT:



```
In [8]: runfile('C:/Users/KIIT/Desktop/Assignments/TNT/Lab2/untitled5.py', wdir='C:/Users/
KIIT/Desktop/Assignments/TNT/Lab2')
   Age  EstimatedSalary  Purchased
0   19            19000          0
1   35            20000          0
2   26            43000          0
3   27            57000          0
4   19            76000          0
```

GRAPH:


Estimated Salary
Age