PANDAS

NUMPY

MATPLOTLIB & SEABORN

In [4]:

```
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv("Social_Network_Ads.csv", sep = ',')
```

In [5]:

```
data.head(10)
```

Out[5]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0
5	27.0	58000.0	0.0
6	27.0	84000.0	0.0
7	32.0	NaN	1.0
8	25.0	33000.0	0.0
9	35.0	65000.0	0.0

In [ ]:

```

```

In [6]:

```
1 data.tail(10)
```

Out[6]:

	Age	EstimatedSalary	Purchased
390	48.0	33000.0	1.0
391	47.0	23000.0	1.0
392	45.0	45000.0	1.0
393	60.0	42000.0	1.0
394	39.0	59000.0	0.0
395	46.0	41000.0	1.0
396	51.0	23000.0	1.0
397	50.0	20000.0	1.0
398	36.0	33000.0	0.0
399	49.0	36000.0	1.0

In [7]:

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Age              398 non-null    float64
 1   EstimatedSalary  397 non-null    float64
 2   Purchased        396 non-null    float64
dtypes: float64(3)
memory usage: 9.5 KB
```

In [8]:

```
1 data.describe()
```

Out[8]:

	Age	EstimatedSalary	Purchased
count	398.000000	397.000000	396.000000
mean	37.680905	69609.571788	0.356061
std	10.492708	33887.085366	0.479439
min	18.000000	15000.000000	0.000000
25%	30.000000	43000.000000	0.000000
50%	37.000000	70000.000000	0.000000
75%	46.000000	87000.000000	1.000000
max	60.000000	150000.000000	1.000000

In [9]:

```
1 data[['Age','EstimatedSalary']]
```

Out[9]:

	Age	EstimatedSalary
0	19.0	19000.0
1	35.0	20000.0
2	26.0	43000.0
3	27.0	57000.0
4	19.0	76000.0
...	...	...
395	46.0	41000.0
396	51.0	23000.0
397	50.0	20000.0
398	36.0	33000.0
399	49.0	36000.0

400 rows × 2 columns

In [10]:

```
1 data.iloc[: , 1:]
```

Out[10]:

	EstimatedSalary	Purchased
0	19000.0	0.0
1	20000.0	0.0
2	43000.0	0.0
3	57000.0	0.0
4	76000.0	0.0
...	...	...
395	41000.0	1.0
396	23000.0	1.0
397	20000.0	1.0
398	33000.0	0.0
399	36000.0	1.0

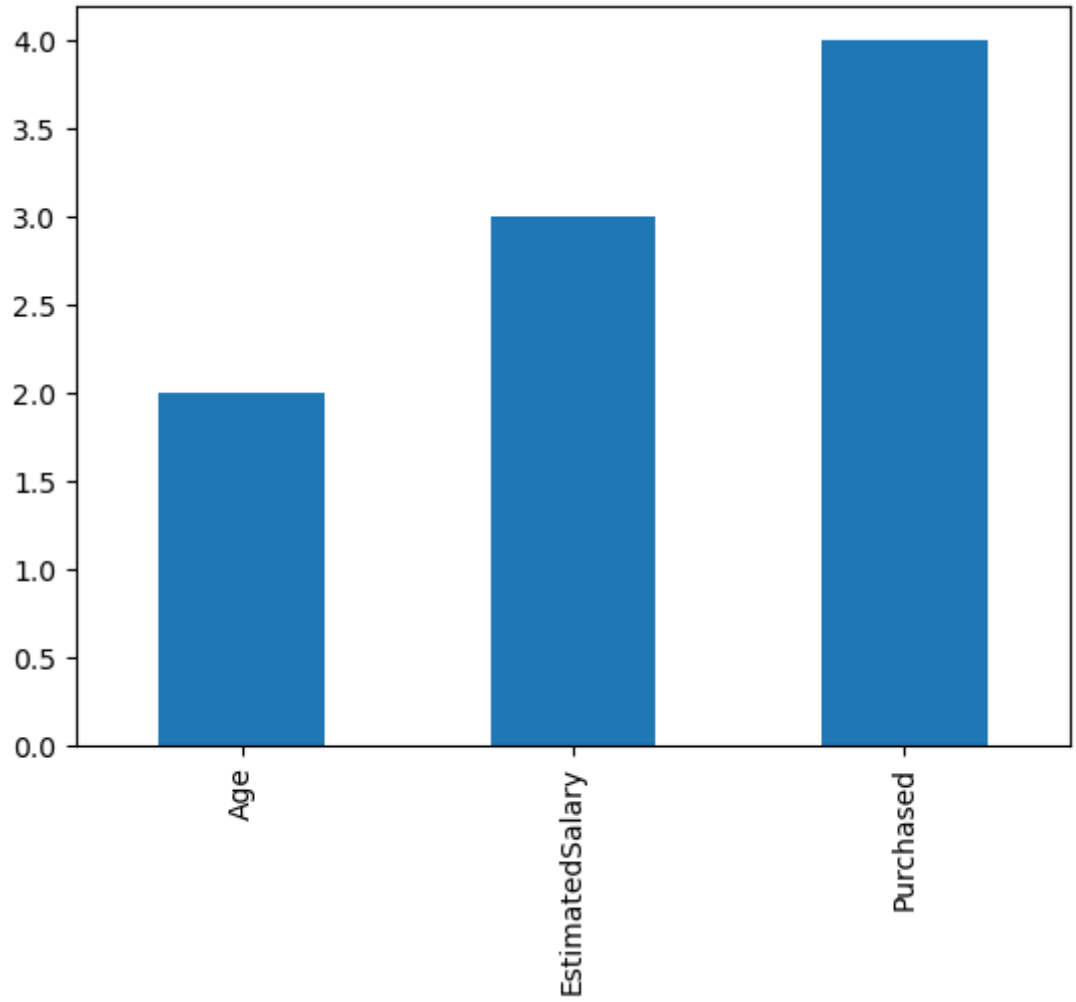
400 rows × 2 columns

In [11]:

```
1 data.isnull().sum().plot(kind = "bar")
```

Out[11]:

<AxesSubplot:>



In [ ]:

```
1
```

In [12]:

```
below_30 = data['Age']<30

data[below_30]
```

Out[12]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0
5	27.0	58000.0	0.0
...	...	...	...
192	29.0	43000.0	0.0
193	19.0	70000.0	0.0
194	28.0	89000.0	0.0
197	20.0	36000.0	0.0
198	26.0	80000.0	0.0

99 rows × 3 columns

In [13]:

```
data[~below_30]
```

Out[13]:

	Age	EstimatedSalary	Purchased
1	35.0	20000.0	0.0
7	32.0	NaN	1.0
9	35.0	65000.0	0.0
13	32.0	18000.0	0.0
16	47.0	25000.0	1.0
...	...	...	...
395	46.0	41000.0	1.0
396	51.0	23000.0	1.0
397	50.0	20000.0	1.0
398	36.0	33000.0	0.0
399	49.0	36000.0	1.0

301 rows × 3 columns

In [ ]:

```
1
```

In [ ]:

```
1
```

In [14]:

```
1 above_30 = data['Age']>30
2 below_30, above_30 = data[~above_30], data[above_30]
3 above_30
```

Out[14]:

	Age	EstimatedSalary	Purchased
1	35.0	20000.0	0.0
7	32.0	NaN	1.0
9	35.0	65000.0	0.0
13	32.0	18000.0	0.0
16	47.0	25000.0	1.0
...	...	...	...
395	46.0	41000.0	1.0
396	51.0	23000.0	1.0
397	50.0	20000.0	1.0
398	36.0	33000.0	0.0
399	49.0	36000.0	1.0

288 rows × 3 columns

In [15]:

```
1 below_30
```

Out[15]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0
5	27.0	58000.0	0.0
...	...	...	...
193	19.0	70000.0	0.0
194	28.0	89000.0	0.0
196	30.0	79000.0	0.0
197	20.0	36000.0	0.0
198	26.0	80000.0	0.0

112 rows × 3 columns

In [16]:

```
1 not_purchased = data['Purchased']==0
2 purchased, not_purchased = data[~not_purchased], data[not_purchased]
3 purchased.describe()
```

Out[16]:

	Age	EstimatedSalary	Purchased
count	145.000000	143.000000	141.0
mean	46.234483	85888.111888	1.0
std	8.655453	41537.714900	0.0
min	27.000000	20000.000000	1.0
25%	39.000000	41500.000000	1.0
50%	47.000000	90000.000000	1.0
75%	53.000000	122000.000000	1.0
max	60.000000	150000.000000	1.0

In [17]:

```
1 not_purchased.describe()
```

Out[17]:

	Age	EstimatedSalary	Purchased
count	253.000000	254.000000	255.0
mean	32.778656	60444.881890	0.0
std	8.025654	24390.493219	0.0
min	18.000000	15000.000000	0.0
25%	26.000000	44000.000000	0.0
50%	34.000000	61000.000000	0.0
75%	38.000000	77000.000000	0.0
max	59.000000	141000.000000	0.0

In [18]:

```
1 data['Purchased'].unique()
```

Out[18]:

```
array([ 0.,  1., nan])
```

In [19]:

```
1 max_val = max(not_purchased["EstimatedSalary"])
2 not_purchased['EstimatedSalary'] == max_val
```

Out[19]:

```
0      False
1      False
2      False
3      False
4      False
       ...
377    False
380    False
387    False
394    False
398    False
Name: EstimatedSalary, Length: 255, dtype: bool
```

In [20]:

```
1 not_purchased[not_purchased['EstimatedSalary'] == max(not_purchased['Estimate
```

Out[20]:

	Age	EstimatedSalary	Purchased
284	48.0	141000.0	0.0

In [21]:

```
1 purchased[purchased['EstimatedSalary'] == min(purchased['EstimatedSalary'].dr
```

Out[21]:

	Age	EstimatedSalary	Purchased
25	47.0	20000.0	1.0
397	50.0	20000.0	1.0

In [ ]:

```
1 data.Age.fillna(data.Age.mean(), inplace = True)
```

In [ ]:

```
1 data.isnull().sum()
```

In [ ]:

```
1
```

In [23]:

```
1 data[['Age']].describe()
```

Out[23]:

	Age
count	398.000000
mean	37.680905
std	10.492708
min	18.000000
25%	30.000000
50%	37.000000
75%	46.000000
max	60.000000

In [24]:

```
1 data.head()
```

Out[24]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0

In [25]:

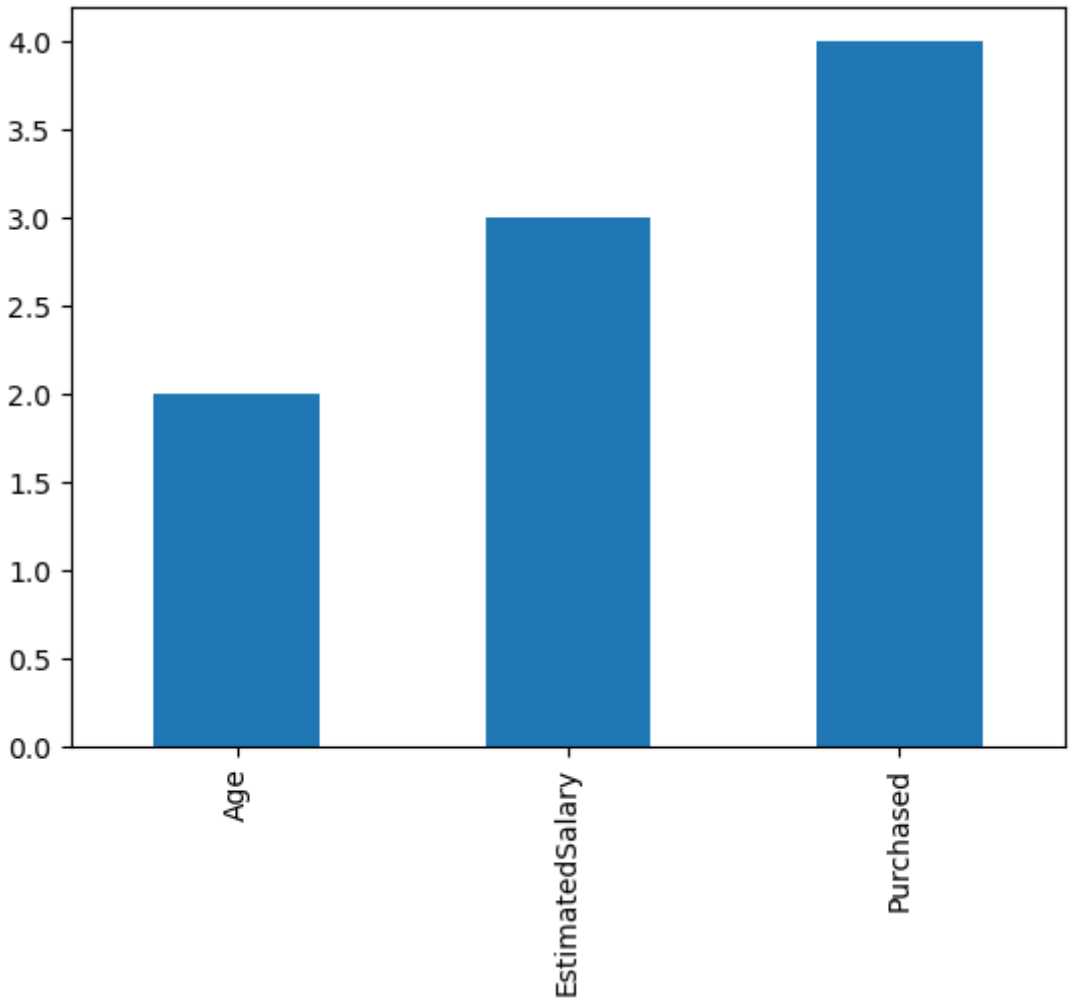
```
1 data.isnull().sum()
```

Out[25]:

```
Age               2
EstimatedSalary   3
Purchased         4
dtype: int64
```

In [26]:

```
data.isnull().sum().plot(kind = 'bar');
```



In [27]:

```
# mising_salary = data['EstimatedSalary'].mean()

# data['EstimatedSalary'].fillna(mising_salary, inplace = True)
```

In [ ]:

```
data.info()
```

In [29]:

```
data.shape
```

Out[29]:

```
(400, 3)
```

In [30]:

```
# df = pd.read_csv('')
```

In [31]:

```
data[data['EstimatedSalary'].isna()]
```

Out[31]:

	Age	EstimatedSalary	Purchased
7	32.0	NaN	1.0
20	45.0	NaN	1.0
34	27.0	NaN	0.0

In [37]:

```
x = data[data['EstimatedSalary'].isna()]
val =[23,45,98]
for i in range(3):
    data[data['EstimatedSalary'].isna()].fillna(val[i], limit = 1)
```

In [38]:

```
x = data[data['EstimatedSalary'].isna()]
x
```

Out[38]:

	Age	EstimatedSalary	Purchased

In [41]:

```
val = [230,450,980]
indexx = [7,20,34]
for i in range(len(val)):
    x.loc[indexx[i], 'EstimatedSalary'] = val[i]
```

```
/tmp/ipykernel_21610/3949083976.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pand
as-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-co
py (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy)
  x.loc[indexx[i], 'EstimatedSalary'] = val[i]
```

In [39]:

```
val = [112,95,11]
indexx = [7,20,34]
for i in range(len(val)):
    data.at[indexx[i], 'EstimatedSalary'] = val[i]
```

In [36]:

```
for i, u in enumerate(val):
    print(f"i_values: {i}\t u_values: ,{u}")
```

```
i_values: 0      u_values: ,112
i_values: 1      u_values: ,95
i_values: 2      u_values: ,11
```

In [42]:

```
for i in range(len(val)):
    print(f"i_values: {i}\t u_values: ,{val[i]}")
```

```
i_values: 0      u_values: ,230
i_values: 1      u_values: ,450
i_values: 2      u_values: ,980
```

In [43]:

```
data[data['EstimatedSalary'].isna()]
```

Out[43]:

Age	EstimatedSalary	Purchased

In [44]:

```
data.head(200)
```

Out[44]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0
...	...	...	...
195	34.0	43000.0	0.0
196	30.0	79000.0	0.0
197	20.0	36000.0	0.0
198	26.0	80000.0	0.0
199	35.0	22000.0	0.0

200 rows × 3 columns

In [45]:

```
1 data[data['Age'].isna()]
```

Out[45]:

	Age	EstimatedSalary	Purchased
108	NaN	86000.0	0.0
110	NaN	71000.0	0.0

In [46]:

```
1 data[data['Purchased'].isna()]
```

Out[46]:

	Age	EstimatedSalary	Purchased
22	48.0	41000.0	NaN
95	35.0	44000.0	NaN
97	28.0	123000.0	NaN
98	35.0	73000.0	NaN

In [47]:

```
1 data[data.isnull().any(axis =1)]
```

Out[47]:

	Age	EstimatedSalary	Purchased
22	48.0	41000.0	NaN
95	35.0	44000.0	NaN
97	28.0	123000.0	NaN
98	35.0	73000.0	NaN
108	NaN	86000.0	0.0
110	NaN	71000.0	0.0

In [48]:

```
1 data.isnull().sum()
```

Out[48]:

```
Age               2
EstimatedSalary   0
Purchased         4
dtype: int64
```

In [49]:

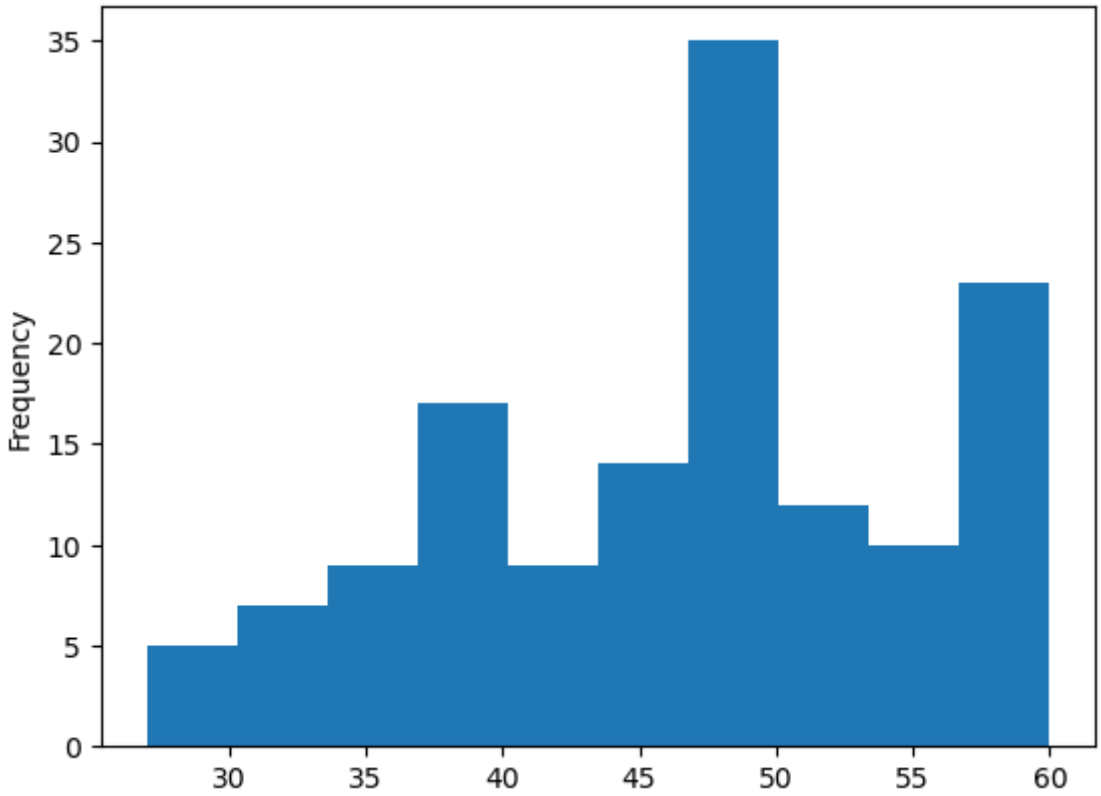
```
1 data.query('(Age < 30) and (Purchased > 0)')
```

Out[49]:

	Age	EstimatedSalary	Purchased
31	27.0	137000.0	1.0
146	27.0	96000.0	1.0
168	29.0	148000.0	1.0

In [50]:

```
1 data.query('(Purchased == 1)').Age.plot(kind= 'hist');
```



In [ ]:

```
1
```

In [51]:

```
1 data['Age'].fillna(data['Age'].mean(), inplace = True)
```

In [52]:

```
1 # np.where(data['EstimatedSalary'] == "")
```

In [53]:

```
1 data.isnull().sum()
```

Out[53]:

```
Age                0
EstimatedSalary    0
Purchased          4
dtype: int64
```

In [54]:

```
1 # data.query("EstimatedSalary == '2'")
```

In [55]:

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Age              400 non-null    float64
 1   EstimatedSalary  400 non-null    float64
 2   Purchased        396 non-null    float64
dtypes: float64(3)
memory usage: 9.5 KB
```

In [56]:

```
1 data.head()
```

Out[56]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0

In [57]:

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Age              400 non-null    float64
 1   EstimatedSalary  400 non-null    float64
 2   Purchased        396 non-null    float64
dtypes: float64(3)
memory usage: 9.5 KB
```

In [58]:

```
1 data
```

Out[58]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0
...	...	...	...
395	46.0	41000.0	1.0
396	51.0	23000.0	1.0
397	50.0	20000.0	1.0
398	36.0	33000.0	0.0
399	49.0	36000.0	1.0

400 rows × 3 columns

creating a bool series for data filtering

In [59]:

```
1 # creating a numpy nan values to be added to a dataframe
2 import numpy as np
3 data_dict = {'age': [1,2,3,4,np.nan,5,np.nan],
4             'width': [np.nan,2,3,4,5,6,np.nan]}
```

In [60]:

```
df = pd.DataFrame.from_dict(data_dict)# creating a dtaframe from dictionary o
```

In [61]:

```
df
```

Out[61]:

	age	width
0	1.0	NaN
1	2.0	2.0
2	3.0	3.0
3	4.0	4.0
4	NaN	5.0
5	5.0	6.0
6	NaN	NaN

In [62]:

```
data
```

Out[62]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0
...	...	...	...
395	46.0	41000.0	1.0
396	51.0	23000.0	1.0
397	50.0	20000.0	1.0
398	36.0	33000.0	0.0
399	49.0	36000.0	1.0

400 rows × 3 columns

In [63]:

```
data_series = pd.notnull(data['Age'])
data_series
```

Out[63]:

```
0      True
1      True
2      True
3      True
4      True
       ...
395    True
396    True
397    True
398    True
399    True
Name: Age, Length: 400, dtype: bool
```

In [64]:

```
data[~data_series]
```

Out[64]:

Age	EstimatedSalary	Purchased

In [65]:

```
data_series = pd.isnull(data['Purchased'])
```

In [66]:

```
data[~data_series]
```

Out[66]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0
...	...	...	...
395	46.0	41000.0	1.0
396	51.0	23000.0	1.0
397	50.0	20000.0	1.0
398	36.0	33000.0	0.0
399	49.0	36000.0	1.0

396 rows × 3 columns

In [67]:

```
1 data[~data_series]
```

Out[67]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0
...	...	...	...
395	46.0	41000.0	1.0
396	51.0	23000.0	1.0
397	50.0	20000.0	1.0
398	36.0	33000.0	0.0
399	49.0	36000.0	1.0

396 rows × 3 columns

In [68]:

```
1 data_series = pd.isnull(data[data.columns])
```

In [69]:

```
1 data[~data_series]
```

Out[69]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0
...	...	...	...
395	46.0	41000.0	1.0
396	51.0	23000.0	1.0
397	50.0	20000.0	1.0
398	36.0	33000.0	0.0
399	49.0	36000.0	1.0

400 rows × 3 columns

In [70]:

```
1 data.head()
```

Out[70]:

	Age	EstimatedSalary	Purchased
0	19.0	19000.0	0.0
1	35.0	20000.0	0.0
2	26.0	43000.0	0.0
3	27.0	57000.0	0.0
4	19.0	76000.0	0.0

In [71]:

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Age              400 non-null    float64
 1   EstimatedSalary  400 non-null    float64
 2   Purchased        396 non-null    float64
dtypes: float64(3)
memory usage: 9.5 KB
```

In [ ]:

```
1
```

In [ ]:

```
1
```