

Pandas

In [1]: `import pandas as pd`

In [2]: `pd.__version__`

Out[2]: '2.2.3'

In [3]: `df = pd.read_csv(r"C:\Users\shaik\OneDrive\Desktop\classroom\17)- data.csv")`
`df`

Out[3]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
1	Afghanistan	AFG	35.253	5.9	Low income
2	Angola	AGO	45.985	19.1	Upper middle income
3	Albania	ALB	12.877	57.2	Upper middle income
4	United Arab Emirates	ARE	11.044	88.0	High income
...
190	Yemen, Rep.	YEM	32.947	20.0	Lower middle income
191	South Africa	ZAF	20.850	46.5	Upper middle income
192	Congo, Dem. Rep.	COD	42.394	2.2	Low income
193	Zambia	ZMB	40.471	15.4	Lower middle income
194	Zimbabwe	ZWE	35.715	18.5	Low income

195 rows × 5 columns

In [4]: `len(df)` *#it will give you records*

Out[4]: 195

In [5]: `id(df)`

Out[5]: 2287468348144

In [6]: `df.shape` *# rows * columns*

Out[6]: (195, 5)

In [7]: `df.columns`

```
Out[7]: Index(['CountryName', 'CountryCode', 'BirthRate', 'InternetUsers',
              'IncomeGroup'],
              dtype='object')
```

```
In [8]: len(df.columns)
```

```
Out[8]: 5
```

```
In [9]: df.isnull()
```

```
Out[9]:
```

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	False	False	False	False	False
1	False	False	False	False	False
2	False	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
...
190	False	False	False	False	False
191	False	False	False	False	False
192	False	False	False	False	False
193	False	False	False	False	False
194	False	False	False	False	False

195 rows × 5 columns

```
In [10]: df.isnull().sum()
```

```
Out[10]: CountryName    0
CountryCode    0
BirthRate    0
InternetUsers    0
IncomeGroup    0
dtype: int64
```

```
In [11]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  -
0   CountryName     195 non-null   object
1   CountryCode     195 non-null   object
2   BirthRate       195 non-null   float64
3   InternetUsers   195 non-null   float64
4   IncomeGroup     195 non-null   object
dtypes: float64(2), object(3)
memory usage: 7.7+ KB
```

```
In [12]: import pandas as pd
```

```
In [13]: df = pd.read_csv(r"C:\Users\shaik\OneDrive\Desktop\classroom\17- data.csv")
```

```
In [14]: df
```

```
Out[14]:
```

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
1	Afghanistan	AFG	35.253	5.9	Low income
2	Angola	AGO	45.985	19.1	Upper middle income
3	Albania	ALB	12.877	57.2	Upper middle income
4	United Arab Emirates	ARE	11.044	88.0	High income
...
190	Yemen, Rep.	YEM	32.947	20.0	Lower middle income
191	South Africa	ZAF	20.850	46.5	Upper middle income
192	Congo, Dem. Rep.	COD	42.394	2.2	Low income
193	Zambia	ZMB	40.471	15.4	Lower middle income
194	Zimbabwe	ZWE	35.715	18.5	Low income

195 rows × 5 columns

```
In [15]: len(df)
```

```
Out[15]: 195
```

```
In [16]: len(df.columns)
```

```
Out[16]: 5
```

```
In [17]: df.shape
```

```
Out[17]: (195, 5)
```

```
In [18]: df
```

Out[18]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
1	Afghanistan	AFG	35.253	5.9	Low income
2	Angola	AGO	45.985	19.1	Upper middle income
3	Albania	ALB	12.877	57.2	Upper middle income
4	United Arab Emirates	ARE	11.044	88.0	High income
...
190	Yemen, Rep.	YEM	32.947	20.0	Lower middle income
191	South Africa	ZAF	20.850	46.5	Upper middle income
192	Congo, Dem. Rep.	COD	42.394	2.2	Low income
193	Zambia	ZMB	40.471	15.4	Lower middle income
194	Zimbabwe	ZWE	35.715	18.5	Low income

195 rows × 5 columns

In [19]: `df.head()`

Out[19]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
1	Afghanistan	AFG	35.253	5.9	Low income
2	Angola	AGO	45.985	19.1	Upper middle income
3	Albania	ALB	12.877	57.2	Upper middle income
4	United Arab Emirates	ARE	11.044	88.0	High income

In [20]: `df.head(2)`

Out[20]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
1	Afghanistan	AFG	35.253	5.9	Low income

In [21]: `df.tail()`

Out[21]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
190	Yemen, Rep.	YEM	32.947	20.0	Lower middle income
191	South Africa	ZAF	20.850	46.5	Upper middle income
192	Congo, Dem. Rep.	COD	42.394	2.2	Low income
193	Zambia	ZMB	40.471	15.4	Lower middle income
194	Zimbabwe	ZWE	35.715	18.5	Low income

In [22]:

df

Out[22]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
1	Afghanistan	AFG	35.253	5.9	Low income
2	Angola	AGO	45.985	19.1	Upper middle income
3	Albania	ALB	12.877	57.2	Upper middle income
4	United Arab Emirates	ARE	11.044	88.0	High income
...
190	Yemen, Rep.	YEM	32.947	20.0	Lower middle income
191	South Africa	ZAF	20.850	46.5	Upper middle income
192	Congo, Dem. Rep.	COD	42.394	2.2	Low income
193	Zambia	ZMB	40.471	15.4	Lower middle income
194	Zimbabwe	ZWE	35.715	18.5	Low income

195 rows × 5 columns

In [23]:

df.dtypes

Out[23]:

```
CountryName    object
CountryCode    object
BirthRate      float64
InternetUsers  float64
IncomeGroup    object
dtype: object
```

In [24]:

df.columns

Out[24]:

```
Index(['CountryName', 'CountryCode', 'BirthRate', 'InternetUsers',
      'IncomeGroup'],
      dtype='object')
```

In [25]: `df['InternetUsers']`

Out[25]:

0	78.9
1	5.9
2	19.1
3	57.2
4	88.0
	...
190	20.0
191	46.5
192	2.2
193	15.4
194	18.5

Name: InternetUsers, Length: 195, dtype: float64

In [26]: `df_cat = df[['CountryName', 'CountryCode', 'IncomeGroup']]`
`df_cat`

Out[26]:

	CountryName	CountryCode	IncomeGroup
0	Aruba	ABW	High income
1	Afghanistan	AFG	Low income
2	Angola	AGO	Upper middle income
3	Albania	ALB	Upper middle income
4	United Arab Emirates	ARE	High income
...
190	Yemen, Rep.	YEM	Lower middle income
191	South Africa	ZAF	Upper middle income
192	Congo, Dem. Rep.	COD	Low income
193	Zambia	ZMB	Lower middle income
194	Zimbabwe	ZWE	Low income

195 rows × 3 columns

In [27]: `df_num = df[['BirthRate', 'InternetUsers']]`
`df_num`

Out[27]:

	BirthRate	InternetUsers
0	10.244	78.9
1	35.253	5.9
2	45.985	19.1
3	12.877	57.2
4	11.044	88.0
...
190	32.947	20.0
191	20.850	46.5
192	42.394	2.2
193	40.471	15.4
194	35.715	18.5

195 rows × 2 columns

```
In [28]: print(df.shape) # 5 columns
print(df_cat.shape) # 3 categorical columns
print(df_num.shape) # 2 numerical columns
```

```
(195, 5)
```

```
(195, 3)
```

```
(195, 2)
```

```
In [29]: df[:]
```

Out[29]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
1	Afghanistan	AFG	35.253	5.9	Low income
2	Angola	AGO	45.985	19.1	Upper middle income
3	Albania	ALB	12.877	57.2	Upper middle income
4	United Arab Emirates	ARE	11.044	88.0	High income
...
190	Yemen, Rep.	YEM	32.947	20.0	Lower middle income
191	South Africa	ZAF	20.850	46.5	Upper middle income
192	Congo, Dem. Rep.	COD	42.394	2.2	Low income
193	Zambia	ZMB	40.471	15.4	Lower middle income
194	Zimbabwe	ZWE	35.715	18.5	Low income

195 rows × 5 columns

In [30]: df[:5]

Out[30]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
1	Afghanistan	AFG	35.253	5.9	Low income
2	Angola	AGO	45.985	19.1	Upper middle income
3	Albania	ALB	12.877	57.2	Upper middle income
4	United Arab Emirates	ARE	11.044	88.0	High income

In [31]: df[5:]

Out[31]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
5	Argentina	ARG	17.716	59.9000	High income
6	Armenia	ARM	13.308	41.9000	Lower middle income
7	Antigua and Barbuda	ATG	16.447	63.4000	High income
8	Australia	AUS	13.200	83.0000	High income
9	Austria	AUT	9.400	80.6188	High income
...
190	Yemen, Rep.	YEM	32.947	20.0000	Lower middle income
191	South Africa	ZAF	20.850	46.5000	Upper middle income
192	Congo, Dem. Rep.	COD	42.394	2.2000	Low income
193	Zambia	ZMB	40.471	15.4000	Lower middle income
194	Zimbabwe	ZWE	35.715	18.5000	Low income

190 rows × 5 columns

In [32]: df[1:200:20]

Out[32]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
1	Afghanistan	AFG	35.253	5.9000	Low income
21	Belize	BLZ	23.092	33.6000	Upper middle income
41	Cuba	CUB	10.400	27.9300	Upper middle income
61	United Kingdom	GBR	12.200	89.8441	High income
81	Ireland	IRL	15.000	78.2477	High income
101	St. Lucia	LCA	15.430	46.2000	Upper middle income
121	Mauritania	MRT	33.801	6.2000	Lower middle income
141	Puerto Rico	PRI	10.800	73.9000	High income
161	Slovak Republic	SVK	10.100	77.8826	High income
181	United States	USA	12.500	84.2000	High income

In [33]: df[:, -1]

Out[33]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
194	Zimbabwe	ZWE	35.715	18.5	Low income
193	Zambia	ZMB	40.471	15.4	Lower middle income
192	Congo, Dem. Rep.	COD	42.394	2.2	Low income
191	South Africa	ZAF	20.850	46.5	Upper middle income
190	Yemen, Rep.	YEM	32.947	20.0	Lower middle income
...
4	United Arab Emirates	ARE	11.044	88.0	High income
3	Albania	ALB	12.877	57.2	Upper middle income
2	Angola	AGO	45.985	19.1	Upper middle income
1	Afghanistan	AFG	35.253	5.9	Low income
0	Aruba	ABW	10.244	78.9	High income

195 rows × 5 columns

In [34]: `df.describe()`

Out[34]:

	BirthRate	InternetUsers
count	195.000000	195.000000
mean	21.469928	42.076471
std	10.605467	29.030788
min	7.900000	0.900000
25%	12.120500	14.520000
50%	19.680000	41.000000
75%	29.759500	66.225000
max	49.661000	96.546800

In [35]: `df.describe().transpose()`

Out[35]:

	count	mean	std	min	25%	50%	75%	max
BirthRate	195.0	21.469928	10.605467	7.9	12.1205	19.68	29.7595	49.6610
InternetUsers	195.0	42.076471	29.030788	0.9	14.5200	41.00	66.2250	96.5468

In [36]: `df_num.describe()`

Out[36]:

	BirthRate	InternetUsers
count	195.000000	195.000000
mean	21.469928	42.076471
std	10.605467	29.030788
min	7.900000	0.900000
25%	12.120500	14.520000
50%	19.680000	41.000000
75%	29.759500	66.225000
max	49.661000	96.546800

In [37]: `df_cat.describe()`

Out[37]:

	CountryName	CountryCode	IncomeGroup
count	195	195	195
unique	195	195	4
top	Aruba	ABW	High income
freq	1	1	67

In [38]: `df.head(1)`

Out[38]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income

In [39]: `df.columns=['a','b','c','d','e']`

In [40]: `df.head(1)`

Out[40]:

	a	b	c	d	e
0	Aruba	ABW	10.244	78.9	High income

In [41]: `df.columns = ['CountryName', 'CountryCode', 'BirthRate', 'InternetUsers', 'IncomeGroup']`

In [42]: `df.head(1)`

Out[42]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income

In [43]: `df.BirthRate * df.InternetUsers`

```
Out[43]: 0      808.2516
         1      207.9927
         2      878.3135
         3      736.5644
         4      971.8720
         ...
        190     658.9400
        191     969.5250
        192      93.2668
        193     623.2534
        194     660.7275
        Length: 195, dtype: float64
```

```
In [44]: df['myCalc'] = df.BirthRate * df.InternetUsers
```

```
In [45]: df.head(1)
```

```
Out[45]:   CountryName  CountryCode  BirthRate  InternetUsers  IncomeGroup  myCalc
0      Aruba          ABW      10.244          78.9    High income    808.2516
```

```
In [46]: df = df.drop('myCalc',axis = 1)
```

```
In [47]: df
```

```
Out[47]:   CountryName  CountryCode  BirthRate  InternetUsers  IncomeGroup
0      Aruba          ABW      10.244          78.9    High income
1  Afghanistan          AFG      35.253           5.9    Low income
2      Angola          AGO      45.985          19.1  Upper middle
   income
3      Albania          ALB      12.877          57.2  Upper middle
   income
4  United Arab          ARE      11.044          88.0    High income
   Emirates
...      ...          ...      ...      ...      ...
190  Yemen, Rep.          YEM      32.947          20.0  Lower middle
   income
191  South Africa          ZAF      20.850          46.5  Upper middle
   income
192  Congo, Dem. Rep.      COD      42.394           2.2    Low income
193      Zambia          ZMB      40.471          15.4  Lower middle
   income
194      Zimbabwe          ZWE      35.715          18.5    Low income
```

195 rows × 5 columns

```
In [48]: df
```

Out[48]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
1	Afghanistan	AFG	35.253	5.9	Low income
2	Angola	AGO	45.985	19.1	Upper middle income
3	Albania	ALB	12.877	57.2	Upper middle income
4	United Arab Emirates	ARE	11.044	88.0	High income
...
190	Yemen, Rep.	YEM	32.947	20.0	Lower middle income
191	South Africa	ZAF	20.850	46.5	Upper middle income
192	Congo, Dem. Rep.	COD	42.394	2.2	Low income
193	Zambia	ZMB	40.471	15.4	Lower middle income
194	Zimbabwe	ZWE	35.715	18.5	Low income

195 rows × 5 columns

In [49]: `df['InternetUsers'] < 2`

Out[49]:

0	False
1	False
2	False
3	False
4	False
...	
190	False
191	False
192	False
193	False
194	False

Name: InternetUsers, Length: 195, dtype: bool

In [50]: `df[df['InternetUsers'] < 2]`

Out[50]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
11	Burundi	BDI	44.151	1.3	Low income
52	Eritrea	ERI	34.800	0.9	Low income
55	Ethiopia	ETH	32.925	1.9	Low income
64	Guinea	GIN	37.337	1.6	Low income
117	Myanmar	MMR	18.119	1.6	Lower middle income
127	Niger	NER	49.661	1.7	Low income
154	Sierra Leone	SLE	36.729	1.7	Low income
156	Somalia	SOM	43.891	1.5	Low income
172	Timor-Leste	TLS	35.755	1.1	Lower middle income

In [51]: `len(df[df['InternetUsers'] < 2])`

Out[51]: 9

In [52]: `filter_1 = df[df['InternetUsers'] < 2]
filter_1`

Out[52]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
11	Burundi	BDI	44.151	1.3	Low income
52	Eritrea	ERI	34.800	0.9	Low income
55	Ethiopia	ETH	32.925	1.9	Low income
64	Guinea	GIN	37.337	1.6	Low income
117	Myanmar	MMR	18.119	1.6	Lower middle income
127	Niger	NER	49.661	1.7	Low income
154	Sierra Leone	SLE	36.729	1.7	Low income
156	Somalia	SOM	43.891	1.5	Low income
172	Timor-Leste	TLS	35.755	1.1	Lower middle income

In [53]: `filter_2 = df[df['BirthRate'] > 40]
filter_2`

Out[53]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
2	Angola	AGO	45.985	19.1	Upper middle income
11	Burundi	BDI	44.151	1.3	Low income
14	Burkina Faso	BFA	40.551	9.1	Low income
65	Gambia, The	GMB	42.525	14.0	Low income
115	Mali	MLI	44.138	3.5	Low income
127	Niger	NER	49.661	1.7	Low income
128	Nigeria	NGA	40.045	38.0	Lower middle income
156	Somalia	SOM	43.891	1.5	Low income
167	Chad	TCD	45.745	2.3	Low income
178	Uganda	UGA	43.474	16.2	Low income
192	Congo, Dem. Rep.	COD	42.394	2.2	Low income
193	Zambia	ZMB	40.471	15.4	Lower middle income

In [54]: `df[(df.BirthRate > 40) & (df.InternetUsers < 2)]`

Out[54]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
11	Burundi	BDI	44.151	1.3	Low income
127	Niger	NER	49.661	1.7	Low income
156	Somalia	SOM	43.891	1.5	Low income

In [55]: `df.head(1)`

Out[55]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income

In [56]: `df[df.IncomeGroup == 'High income'].head()`

Out[56]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.9	High income
4	United Arab Emirates	ARE	11.044	88.0	High income
5	Argentina	ARG	17.716	59.9	High income
7	Antigua and Barbuda	ATG	16.447	63.4	High income
8	Australia	AUS	13.200	83.0	High income

In [57]: `df.IncomeGroup.unique()`Out[57]: `array(['High income', 'Low income', 'Upper middle income', 'Lower middle income'], dtype=object)`

```
In [58]: df.IncomeGroup.nunique()
```

```
Out[58]: 4
```

```
In [59]: import matplotlib.pyplot as plt
import seaborn as sns

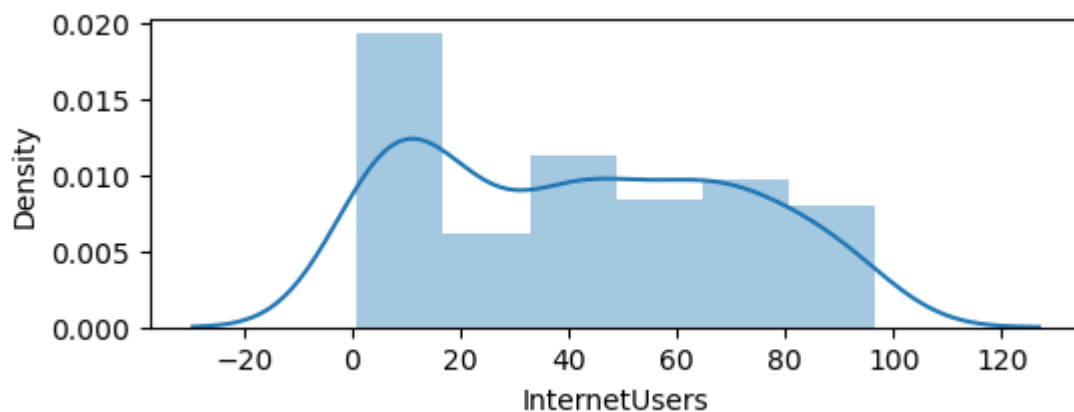
%matplotlib inline
plt.rcParams['figure.figsize'] = 6,2

import warnings
warnings.filterwarnings('ignore')
```

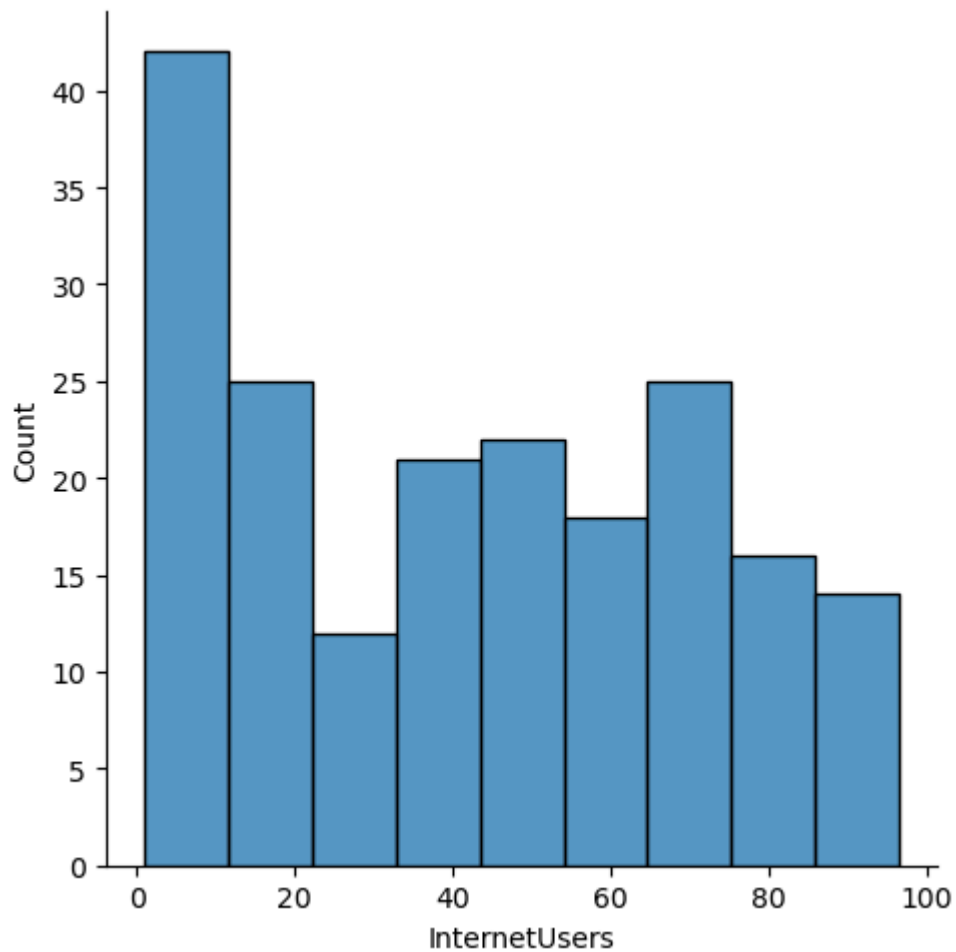
```
In [60]: df["InternetUsers"]
```

```
Out[60]: 0      78.9
1       5.9
2      19.1
3      57.2
4      88.0
...
190    20.0
191    46.5
192     2.2
193    15.4
194    18.5
Name: InternetUsers, Length: 195, dtype: float64
```

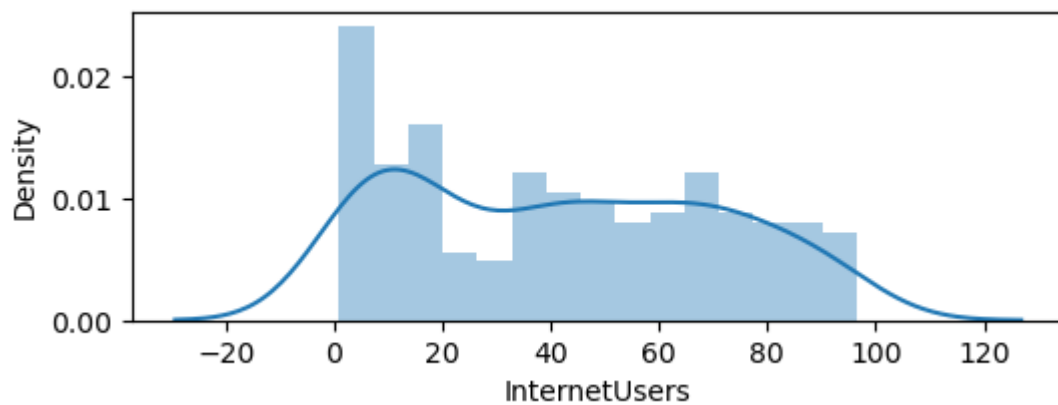
```
In [61]: vis1 = sns.distplot(df["InternetUsers"])
plt.show()
```



```
In [69]: vis2 = sns.displot(df["InternetUsers"])
plt.show()
```

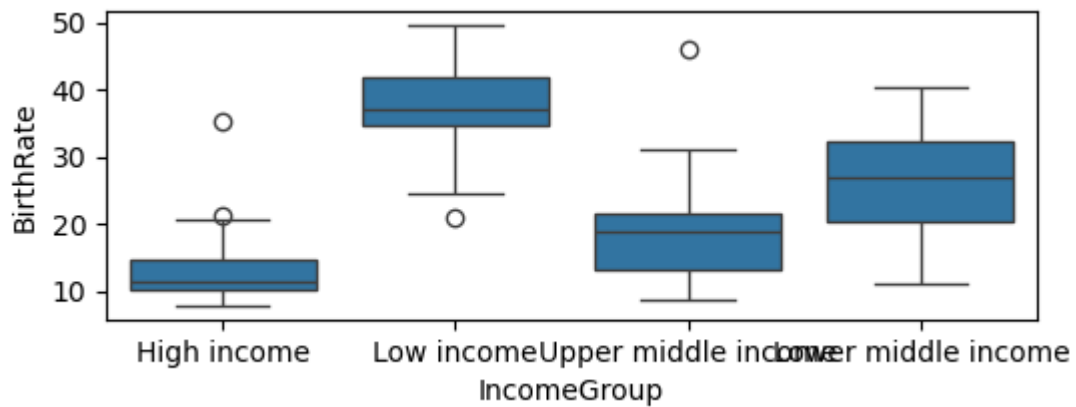



```
In [70]: vis3 = sns.distplot(df["InternetUsers"],bins=15)  
plt.show()
```



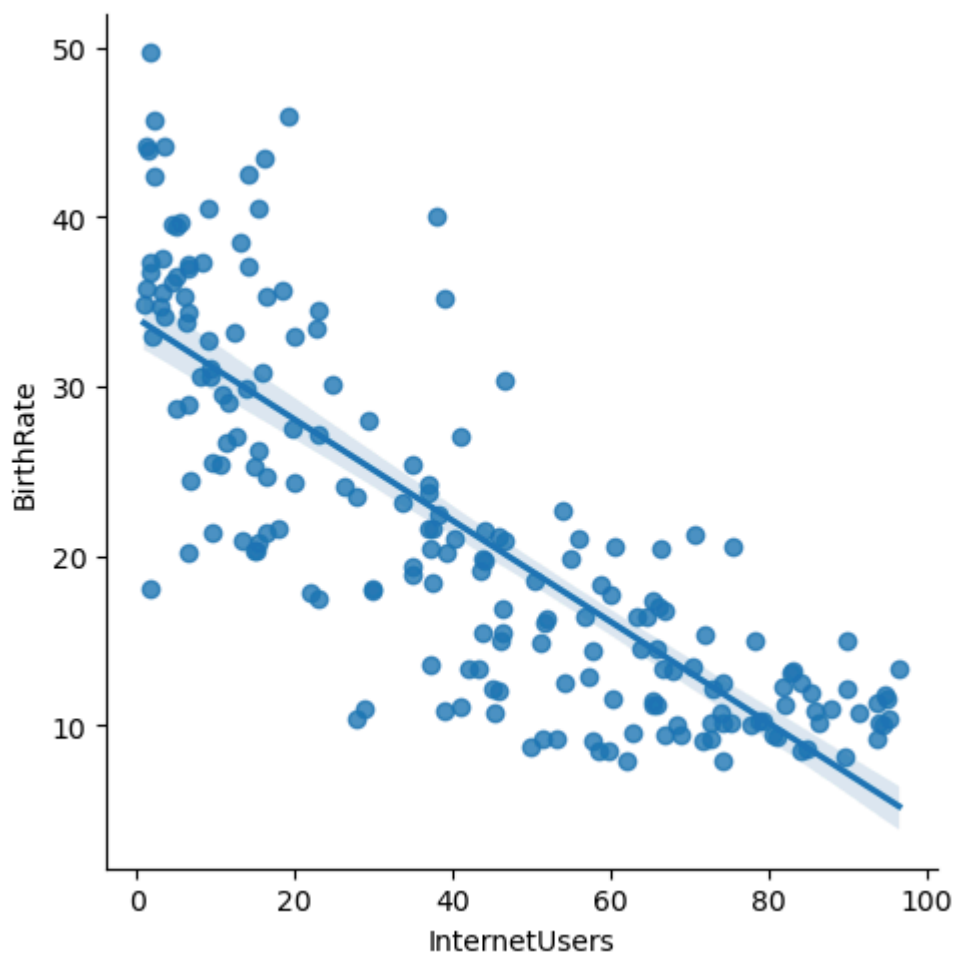
```
In [71]: # BoxPlot:
```

```
In [72]: vis4=sns.boxplot(data=df,x="IncomeGroup",y="BirthRate")  
plt.show()
```

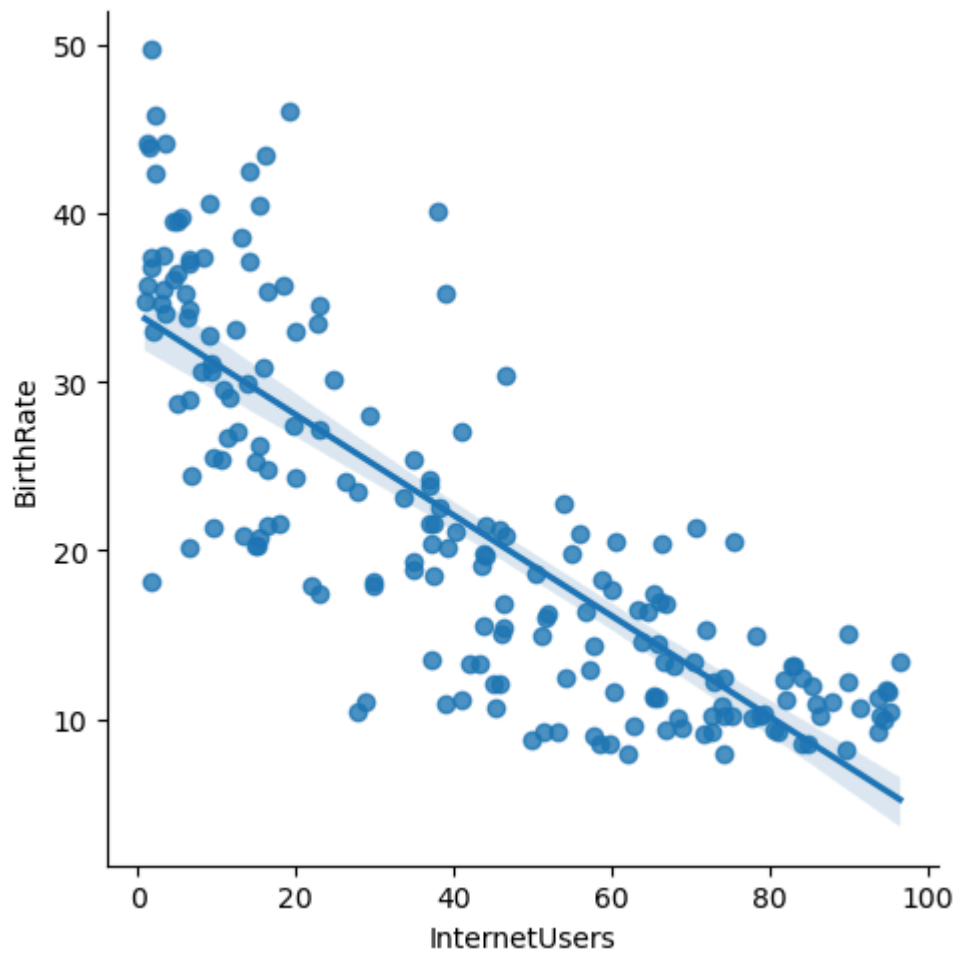


```
In [73]: # outliers also called as ANOMALY DTECTION
```

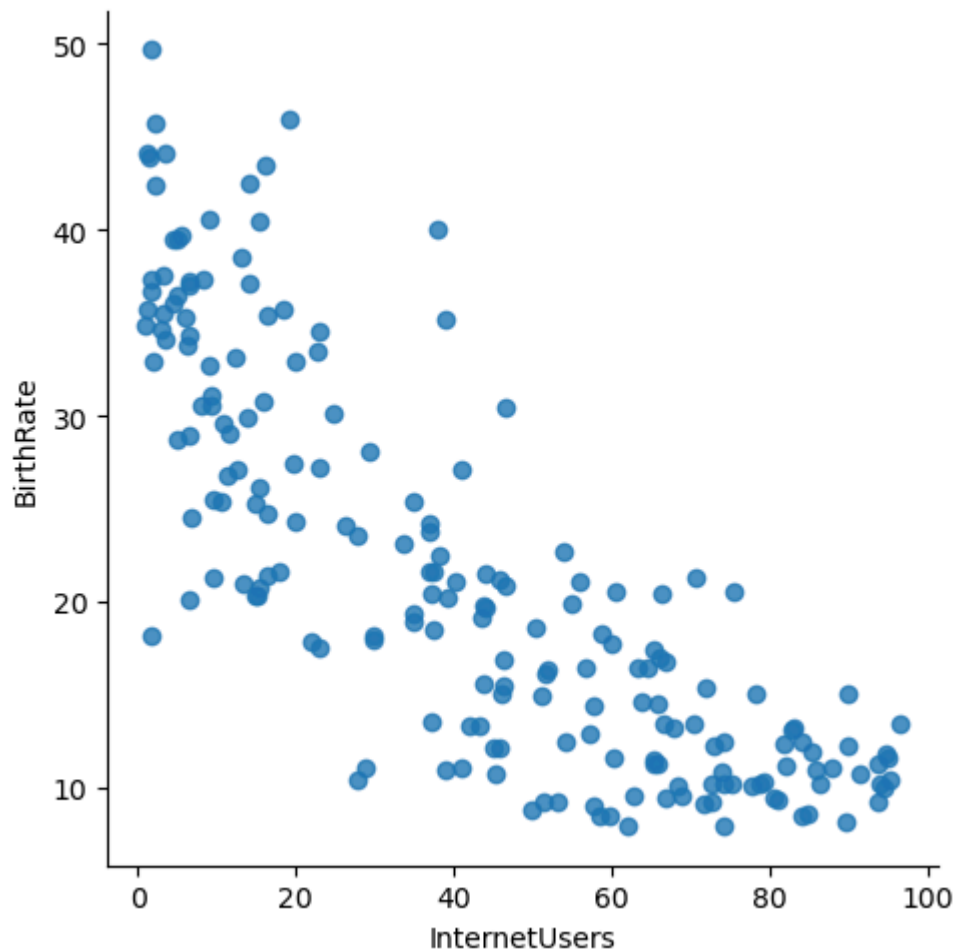
```
In [74]: vis5=sns.lmplot(data=df,x="InternetUsers",y="BirthRate")
plt.show()
```



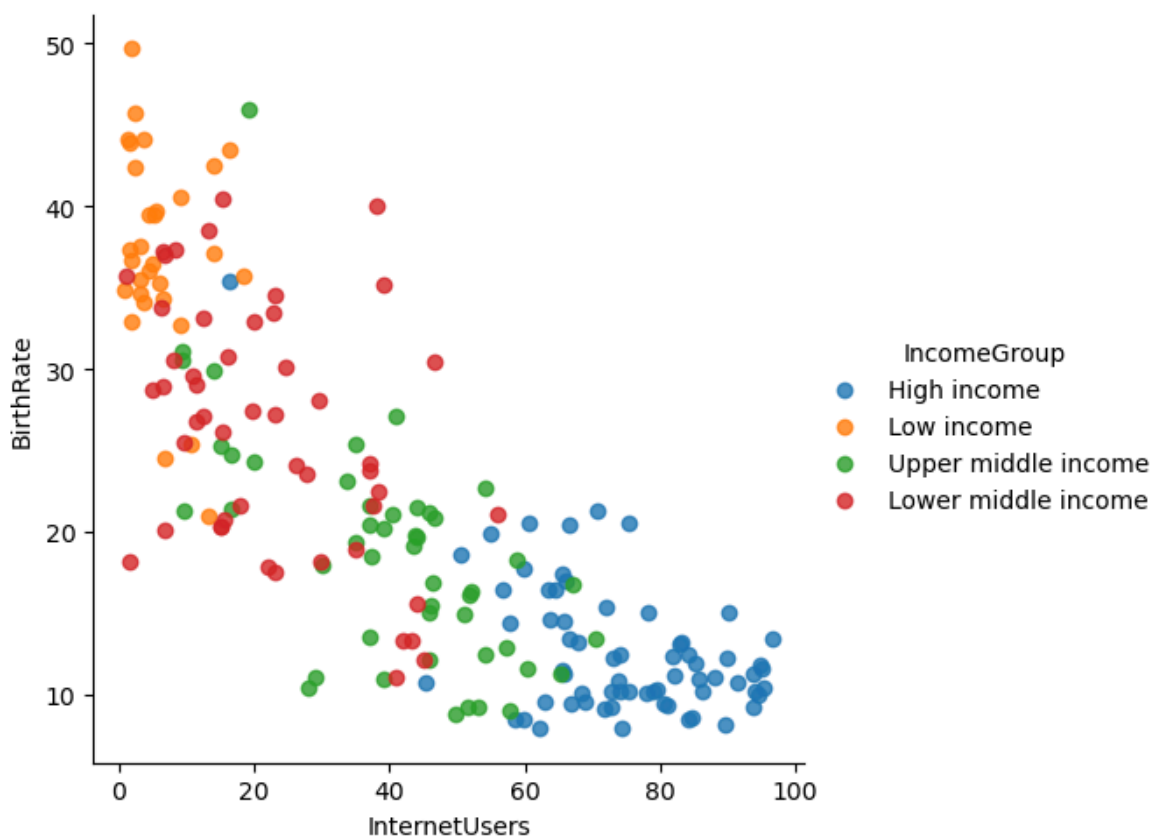
```
In [75]: vis5=sns.lmplot(data=df,x="InternetUsers",y="BirthRate",fit_reg=True)
plt.show()
```



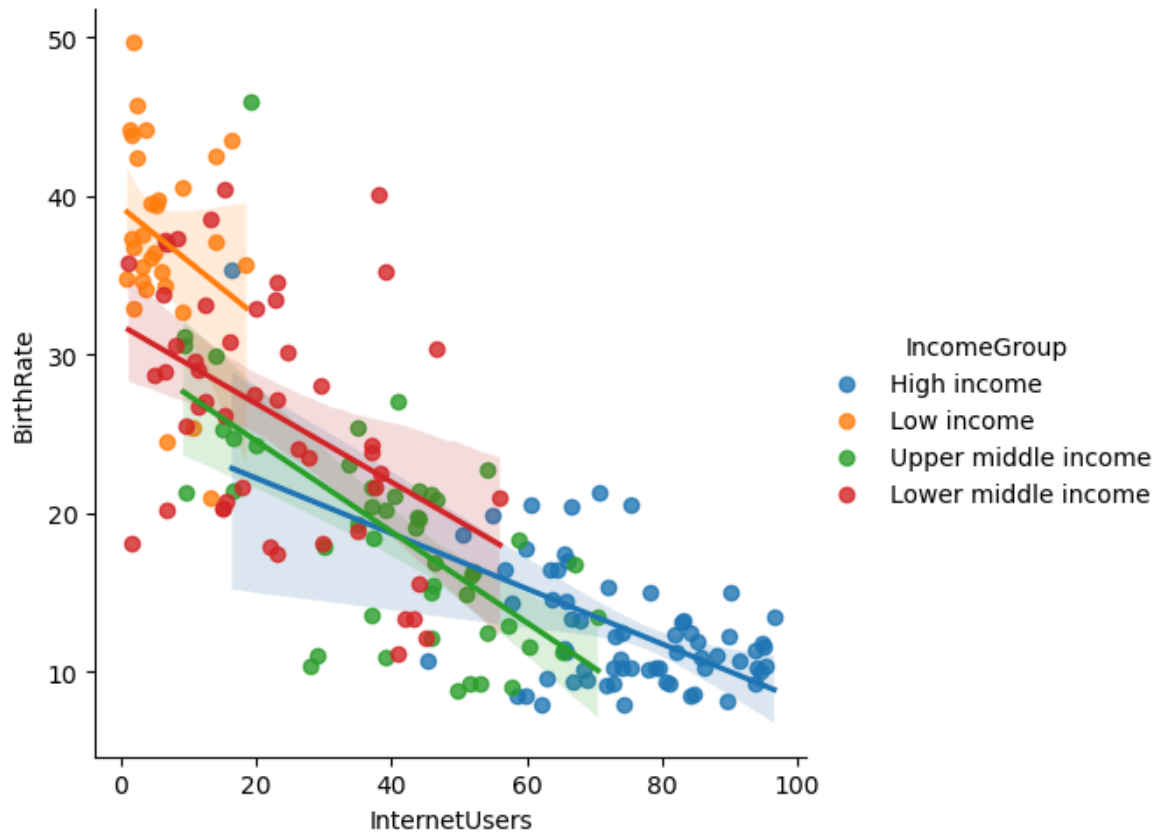
```
In [76]: vis5=sns.lmplot(data=df,x="InternetUsers",y="BirthRate",fit_reg=False)
plt.show()
```



```
In [80]: vis6=sns.lmplot(data=df,x="InternetUsers",y="BirthRate",fit_reg=False,hue="IncomeGroup",plt.show())
```



```
In [82]: vis7=sns.lmplot(data=df,x="InternetUsers",y="BirthRate",fit_reg=True,hue="IncomeGroup",plt.show())
```



```
In [ ]:
```