In [2]:

```
import pandas as pd
import numpy as np
```

In [4]:

```
df = pd.read_csv(r'C:\Users\LENOVO\Desktop\Monty Datascien\1st, 2nd\DataFrame_ Pandas\DemographicData.csv')
```

In [5]:

```
df
```

Out[5]:

	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 [6]:

```
df.shape
```

Out[6]:

```
(195, 5)
```

In [7]:

```
df.size
```

Out[7]:

```
975
```

In [8]:

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

Out[8]:

```
CountryName      0
CountryCode      0
BirthRate        0
InternetUsers    0
IncomeGroup      0
dtype: int64
```

In [10]:

```
df.columns # for getting each attributes/features/independent variables of dataset.
```

Out[10]:

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

In [11]:

```
df.head()  # for getting top 5 entries from dataset
```

Out[11]:

	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 [12]:

```
df.tail()   # for getting last 5 entries from dataset
```

Out[12]:

	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 [13]:

```
len(df)
```

Out[13]:

```
195
```

In [15]:

```
import matplotlib.pyplot as plt # used for normal visualization
import seaborn as sns          # used for advance visualization

%matplotlib inline
plt.rcParams['figure.figsize'] = 10,5

import warnings
warnings.filterwarnings('ignore')   # use for not getting os error
```

In [16]:

```
# for descriptive statistics - we only get numerical data
df.describe()
```

Out[16]:

	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 [17]:

```
# for finding information about dataset
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 [18]:

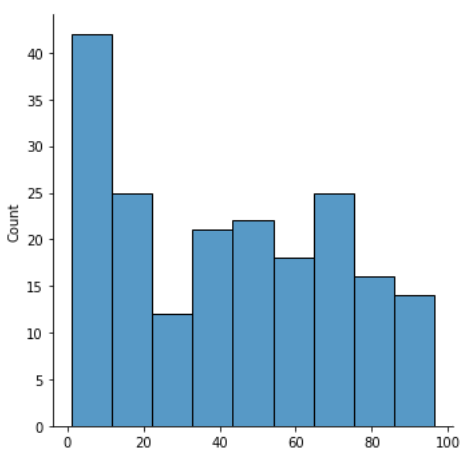
```
df['InternetUsers']
```

Out[18]:

```
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
```

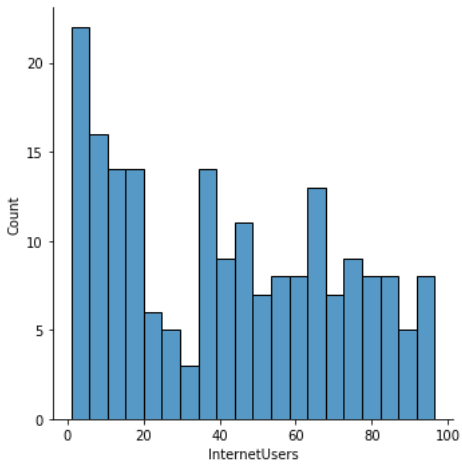
In [21]:

```
v = sns.displot(df['InternetUsers'],kind='hist') # this is univariate analysis because we are using only one variable
```



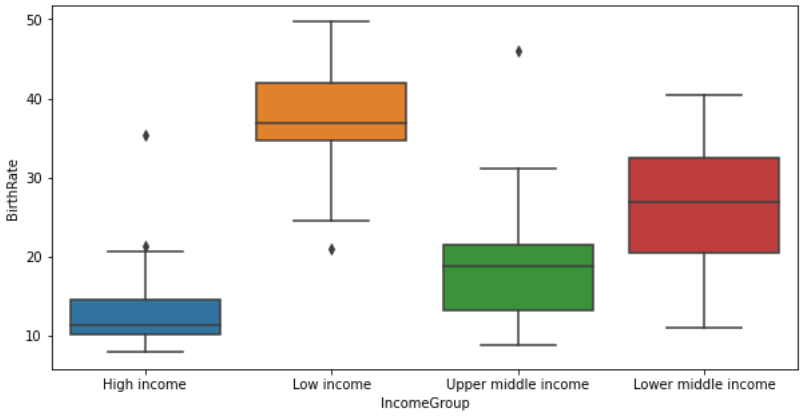
In [20]:

```
v = sns.displot(df['InternetUsers'],bins=20)
```



In [24]:

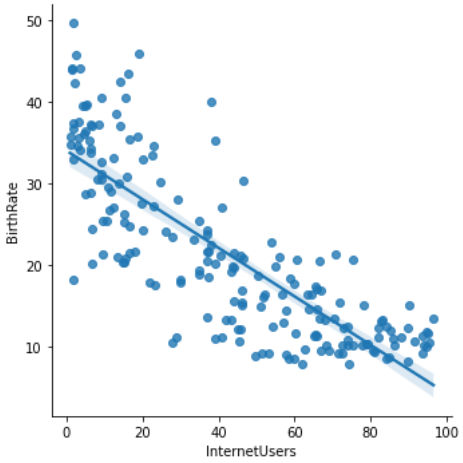
```
v1 = sns.boxplot(data=df, x= 'IncomeGroup', y ='BirthRate') # Univariate Analysis
```



Here we have outlier values in High income(1),Low income(1),upper middle income(1) , totally 3//.. so we have to ask our project manager how to handle them(means should we simply remove outlier or do modifications)

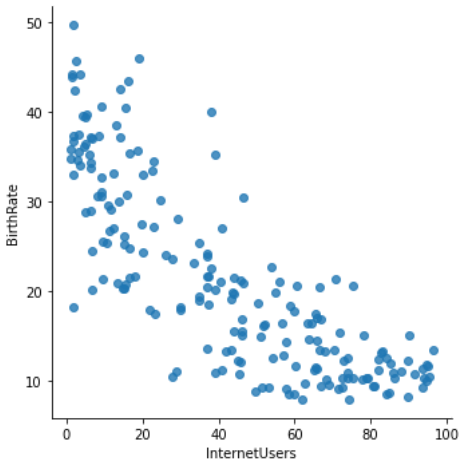
In [25]:

```
v2 = sns.lmplot(data=df, x='InternetUsers', y='BirthRate')
```



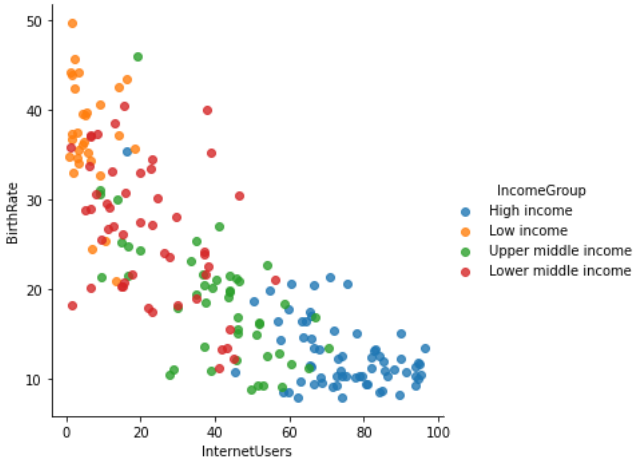
In [26]:

```
v2 = sns.lmplot(data=df, x='InternetUsers', y='BirthRate', fit_reg=False) # we will not get linear line in the graph if fit_reg=False
```



In [27]:

```
'BirthRate', fit_reg=False, hue ='IncomeGroup') # hue - it is used to give different color for parameters so that we can do the analysis.
```



what understood from visualization

1. Low income group have high birthrate compared with others in our dataset?why?
2. Low income group have less number of internet users?why?
3. High income group have less birthrate and high internet users?why?
4. Upper Middle income group have high internetuser and less birthrate and viceversa for lower middle income
5. Is there any relationship b/w Internetuers and birthrate(As we have seen less internetusers have more birthrate)

In [ ]: