

```
In [54]: import pandas as pd
pd.__version__
```

Out[54]: '2.2.2'

```
In [55]: df = pd.read_csv('data.csv')
df
```

Out[55]:

	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 [56]: id(df)
```

Out[56]: 1983617040064

```
In [57]: df.head()
```

Out[57]:

	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 [58]: df.tail()
```

Out[58]:

	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 [59]:

len(df)

Out[59]: 195

In [60]:

df.columns

Out[60]: Index(['CountryName', 'CountryCode', 'BirthRate', 'InternetUsers', 'IncomeGroup'], dtype='object')

In [61]:

df.isnull()

Out[61]:

	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 [62]:

df.isna()

Out[62]:

	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 [63]:

```
# counting missing values in each column
df.isnull().sum()
```

Out[63]:

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

In [64]:

```
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 [65]:

```
df[1:11]
```

Out[65]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
1	Afghanistan	AFG	35.253	5.9000	Low income
2	Angola	AGO	45.985	19.1000	Upper middle income
3	Albania	ALB	12.877	57.2000	Upper middle income
4	United Arab Emirates	ARE	11.044	88.0000	High income
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
10	Azerbaijan	AZE	18.300	58.7000	Upper middle income

In [66]: df[1:100:10]

Out[66]:

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
1	Afghanistan	AFG	35.253	5.9000	Low income
11	Burundi	BDI	44.151	1.3000	Low income
21	Belize	BLZ	23.092	33.6000	Upper middle income
31	Switzerland	CHE	10.200	86.3400	High income
41	Cuba	CUB	10.400	27.9300	Upper middle income
51	Egypt, Arab Rep.	EGY	28.032	29.4000	Lower middle income
61	United Kingdom	GBR	12.200	89.8441	High income
71	Guatemala	GTM	27.465	19.7000	Lower middle income
81	Ireland	IRL	15.000	78.2477	High income
91	Kenya	KEN	35.194	39.0000	Lower middle income

In [67]: df.head(4)

Out[67]:

	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

```
In [68]: df.describe()
```

```
Out[68]:      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 [69]: df['CountryName']
```

```
Out[69]: 0           Aruba
1           Afghanistan
2           Angola
3           Albania
4           United Arab Emirates
...
190          Yemen, Rep.
191          South Africa
192          Congo, Dem. Rep.
193          Zambia
194          Zimbabwe
Name: CountryName, Length: 195, dtype: object
```

```
In [70]: df['CountryCode']
```

```
Out[70]: 0      ABW
1      AFG
2      AGO
3      ALB
4      ARE
...
190    YEM
191    ZAF
192    COD
193    ZMB
194    ZWE
Name: CountryCode, Length: 195, dtype: object
```

```
In [71]: df[['CountryName', 'CountryCode']]
```

Out[71]:

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

195 rows × 2 columns

In [72]:

```
print(df.columns)
```

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

In [73]:

```
df_cat = df[['CountryName', 'CountryCode', 'IncomeGroup']]
```

In [74]:

```
df_cat
```

Out[74]:

	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 [75]:

df_cat.describe()

Out[75]:

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

In [76]:

df_num = df[['BirthRate', 'InternetUsers']]

In [77]:

df_num

Out[77]:

	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 [78]:

df_num.describe()

Out[78]:

	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 [79]:

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 [80]: `df_cat.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   CountryName    195 non-null   object  
 1   CountryCode     195 non-null   object  
 2   IncomeGroup     195 non-null   object  
dtypes: object(3)
memory usage: 4.7+ KB
```

In [81]: `df_num.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   BirthRate      195 non-null   float64 
 1   InternetUsers  195 non-null   float64 
dtypes: float64(2)
memory usage: 3.2 KB
```

In [82]: `df.describe().transpose()`

	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 [83]: `# shortcut
df.describe().T`

	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 [84]: # Renaming column
df.columns = ['a','b','c','d','e']
```

```
In [85]: df.head()
```

Out[85]:

	a	b	c	d	e
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 [86]: df.columns = ['CountryName','CountryCode ','BirthRate', 'InternetUsers', 'IncomeGroup']
```

```
In [87]: df.head(1)
```

Out[87]:

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

```
In [88]: df[['CountryName','BirthRate','InternetUsers']][4:8] #subset
```

Out[88]:

	CountryName	BirthRate	InternetUsers
4	United Arab Emirates	11.044	88.0
5	Argentina	17.716	59.9
6	Armenia	13.308	41.9
7	Antigua and Barbuda	16.447	63.4

```
In [89]: df[4:8][['CountryName','BirthRate','InternetUsers']]
```

Out[89]:

	CountryName	BirthRate	InternetUsers
4	United Arab Emirates	11.044	88.0
5	Argentina	17.716	59.9
6	Armenia	13.308	41.9
7	Antigua and Barbuda	16.447	63.4

```
In [90]: df.BirthRate * df.InternetUsers
```

```
Out[90]: 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 [91]: # Adding new column in df
df['Newcolumn'] = df.BirthRate * df.InternetUsers
```

```
In [92]: df.head(1)
```

```
Out[92]:   CountryName  CountryCode  BirthRate  InternetUsers  IncomeGroup  Newcolumn
          0           Aruba       ABW      10.244        78.9    High income    808.2516
```

```
In [93]: len(df.columns)
```

```
Out[93]: 6
```

```
In [94]: df = df.drop('Newcolumn', axis=1)
```

```
In [95]: len(df.columns)
```

```
Out[95]: 5
```

```
In [96]: df
```

Out[96]:

	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 [97]: df.InternetUsers < 2

```
Out[97]: 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 [98]: df[df.InternetUsers < 2]

Out[98]:

	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 [99]:

len(df[df.InternetUsers < 2])

Out[99]:

9

In [100...]

df[df.BirthRate > 40]

Out[100...]

	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 [101...]

len(df[df.BirthRate > 40])

Out[101...]

12

In [102...]

filter = df.InternetUsers < 2

```
In [103... filter2 = df.BirthRate > 40
```

```
In [104... df[filter & filter2]
```

	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

11th July

```
In [109... df[df.IncomeGroup == 'High income']
```

	CountryName	CountryCode	BirthRate	InternetUsers	IncomeGroup
0	Aruba	ABW	10.244	78.90	High income
4	United Arab Emirates	ARE	11.044	88.00	High income
5	Argentina	ARG	17.716	59.90	High income
7	Antigua and Barbuda	ATG	16.447	63.40	High income
8	Australia	AUS	13.200	83.00	High income
...
174	Trinidad and Tobago	TTO	14.590	63.80	High income
180	Uruguay	URY	14.374	57.69	High income
181	United States	USA	12.500	84.20	High income
184	Venezuela, RB	VEN	19.842	54.90	High income
185	Virgin Islands (U.S.)	VIR	10.700	45.30	High income

67 rows × 5 columns

Printing unique values

```
In [106... df.IncomeGroup.unique()
```

```
Out[106... array(['High income', 'Low income', 'Upper middle income',
       'Lower middle income'], dtype=object)
```

Number of Unique values

```
In [107... df.IncomeGroup.unique()
```

```
Out[107... 4
```

```
In [112... import matplotlib.pyplot as plt # visualizations
import seaborn as sns # stats and advance visualization
%matplotlib inline # plot the graph in the line
plt.rcParams['figure.figsize'] = 6,2 #

import warnings
warnings.filterwarnings('ignore') # os error
```

```
In [113... df.head()
```

	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 [114... df.columns
```

```
Out[114... Index(['CountryName', 'CountryCode ', 'BirthRate', 'InternetUsers',
       'IncomeGroup'],
       dtype='object')
```

```
In [115... df.InternetUsers
```

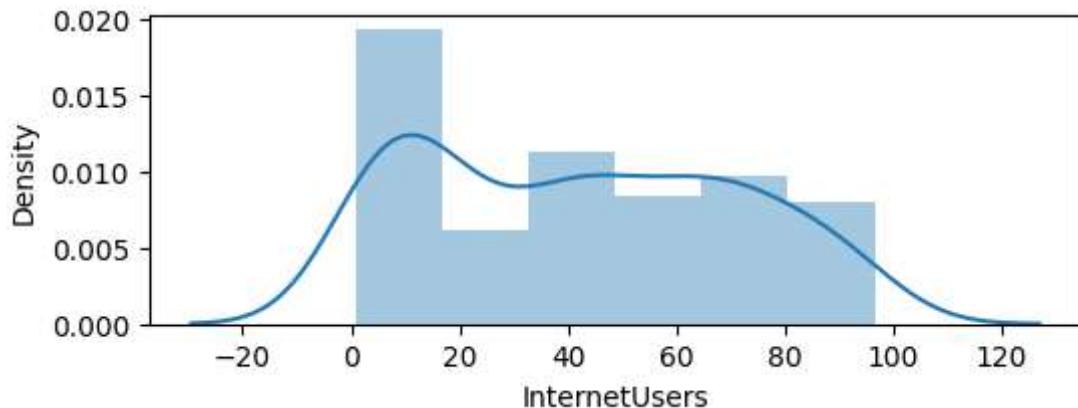
```
Out[115... 0    78.9
1    5.9
2    19.1
3    57.2
4    88.0
...
190   20.0
191   46.5
192    2.2
193   15.4
194   18.5
Name: InternetUsers, Length: 195, dtype: float64
```

Univariate analysis

(Plot the graph using 1 variable is called univariate analysis)

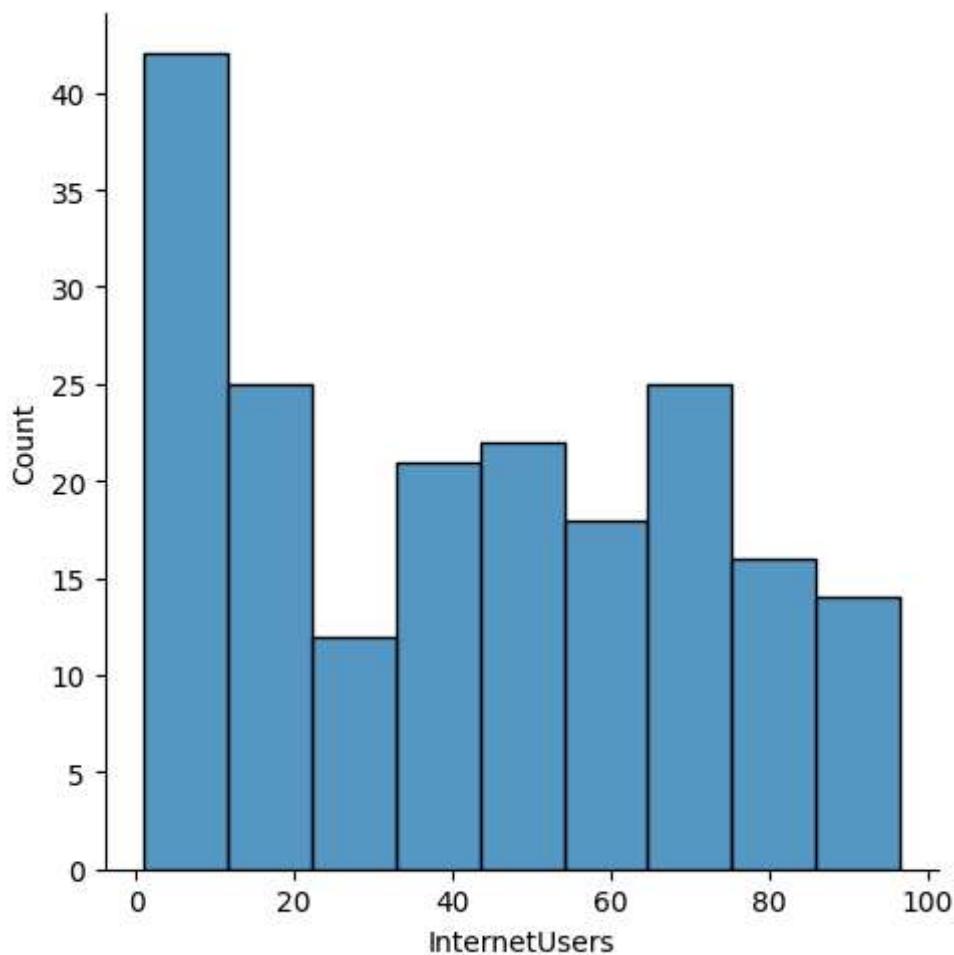
In [117...]

```
vis1 = sns.distplot(df['InternetUsers'])
```



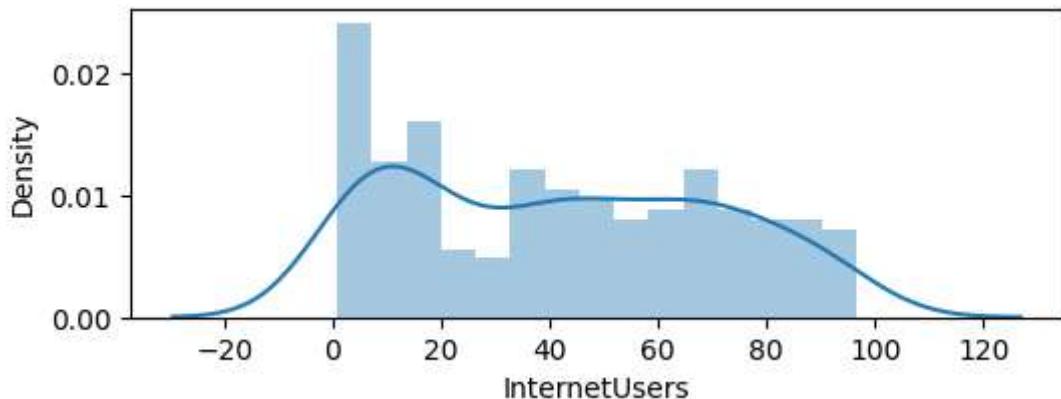
In [119...]

```
vis2 = sns.distplot(df['InternetUsers'])
```



In [120...]

```
vis3 = sns.distplot(df['InternetUsers'], bins=15) # adding more bars for more indept
```

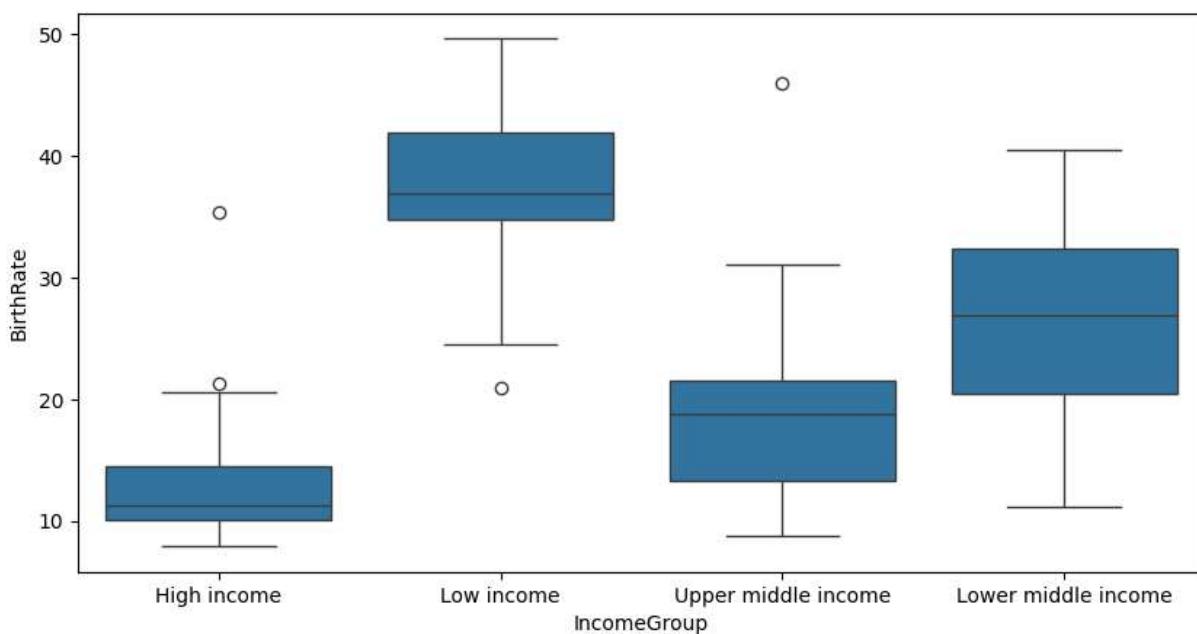


Bivariate Analysis

Plot graph using two variables is called Bivariate Analysis

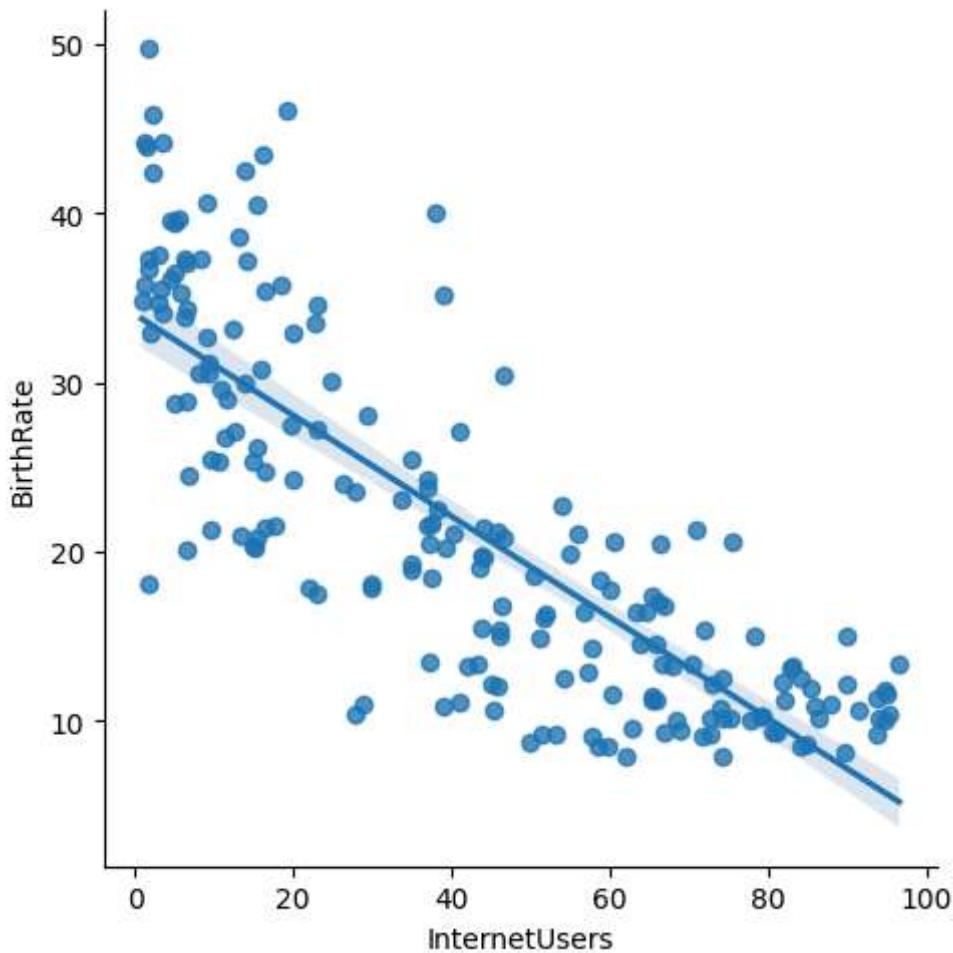
In [128...]

```
plt.rcParams['figure.figsize'] = 10,5  
vis4 = sns.boxplot(data = df, x = 'IncomeGroup', y = 'BirthRate')
```

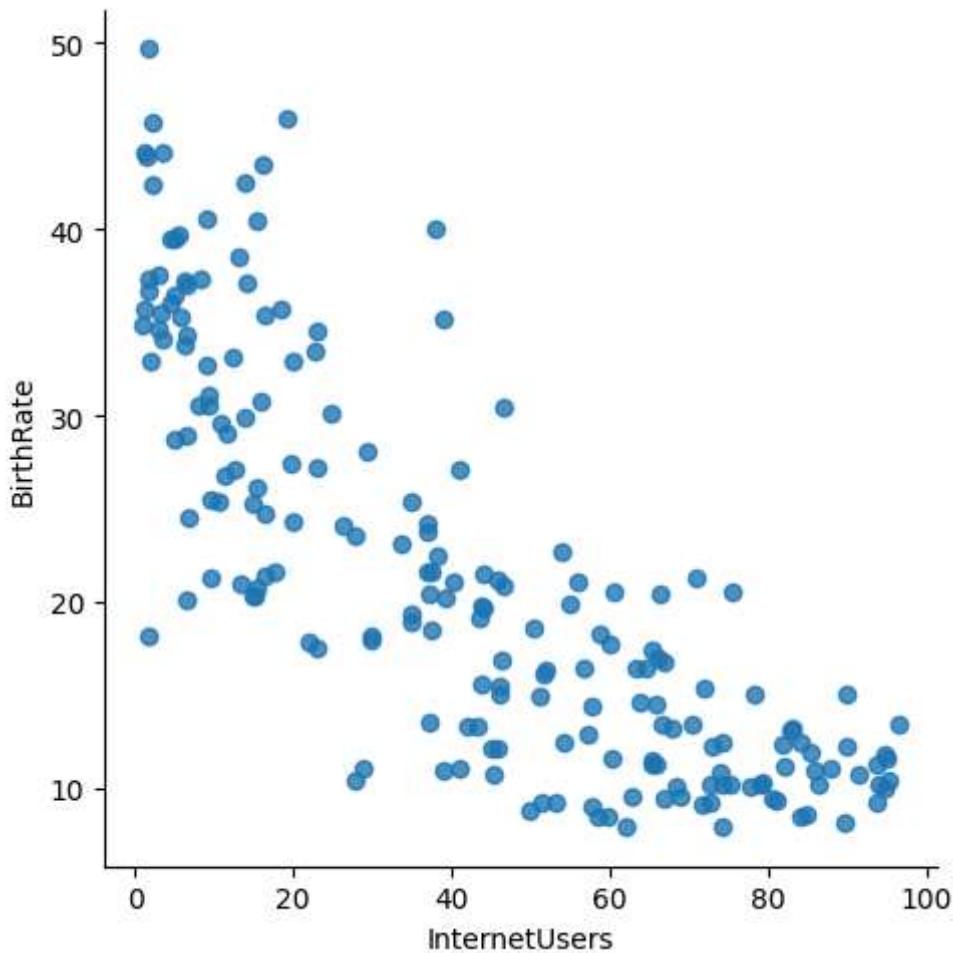


In [130...]

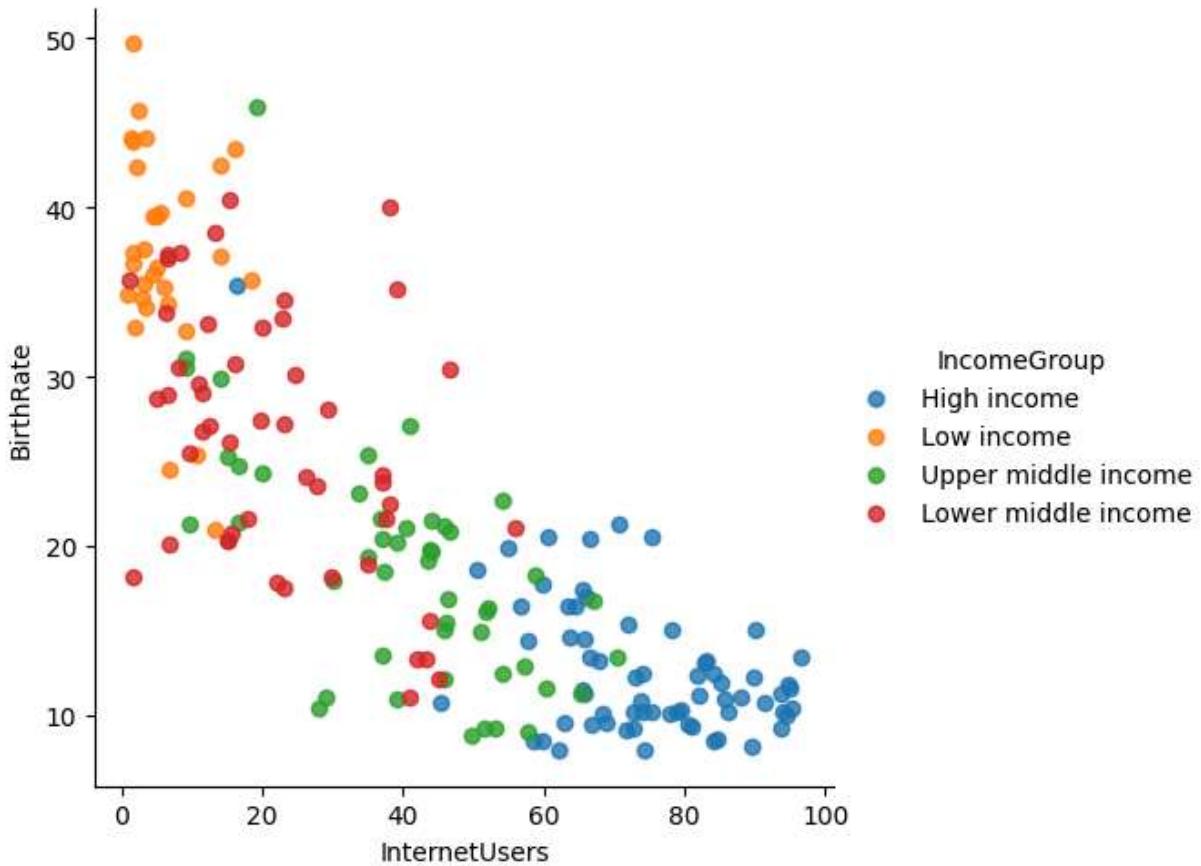
```
vis5 = sns.lmplot(data = df, x = 'InternetUsers', y = 'BirthRate')
```



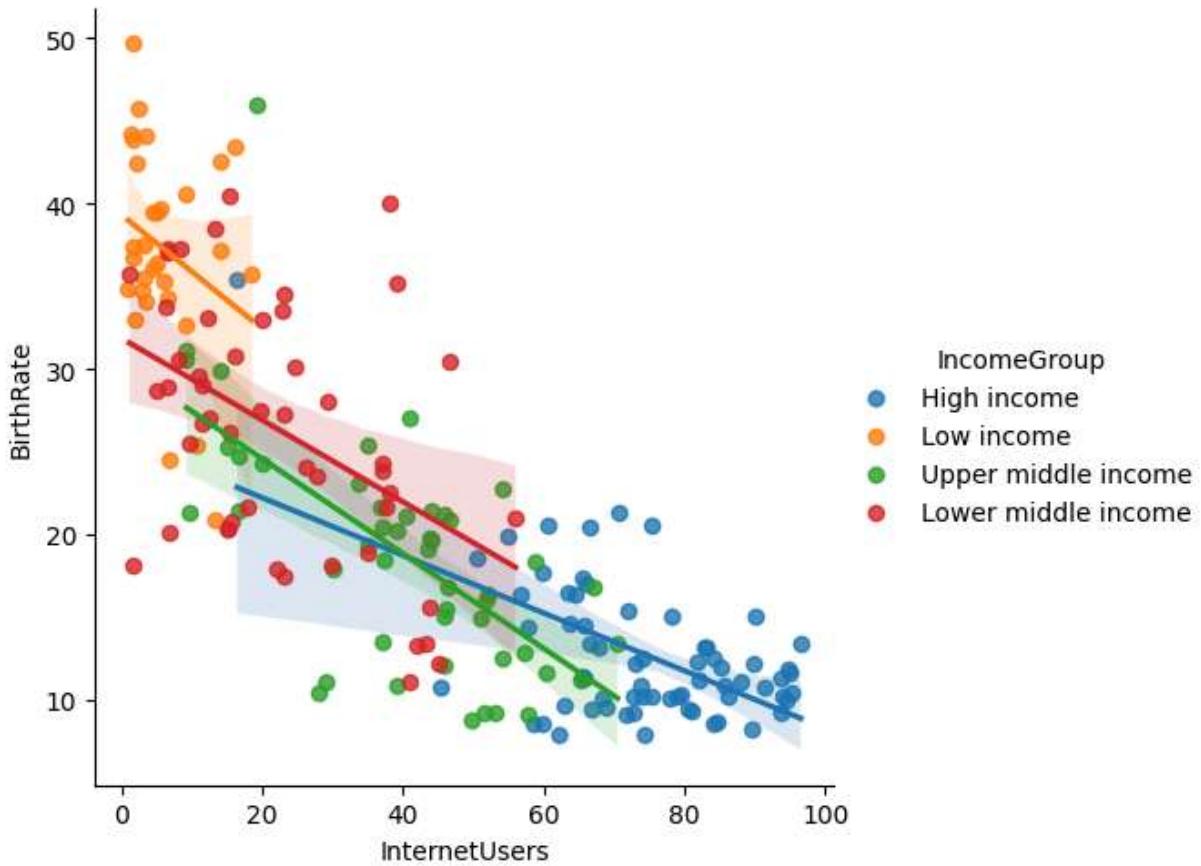
```
In [131]: vis5 = sns.lmplot(data = df, x = 'InternetUsers', y = 'BirthRate', fit_reg=False)
```



```
In [132]: vis5 = sns.lmplot(data = df, x = 'InternetUsers', y = 'BirthRate', fit_reg = False,
```



```
In [133]: vis5 = sns.lmplot(data = df, x = 'InternetUsers', y = 'BirthRate', fit_reg = True,
```



In []: