DataFrames

- A Data frame is a two-dimensional data structure
- Data is aligned in a tabular fashion in rows and columns

```
import pandas as pd
statsDF = pd.read_csv('C:\\.....\\file1.csv')
```

DataFrames

- A Data frame is a two-dimensional data structure
- Data is aligned in a tabular fashion in rows and columns

```
import pandas as pd
statsDF = pd.read_csv('C:\\.....\\file1.csv')
```

Exploring DataSet

len(df)

df.head())

len(df.columns)

df.columns

df.tail()

df.info()

df.describe()

Rename Columns

Standard Deviation

Customer ID	Name	Surname	Gender	Age	Age Grou	Height	egion	Job Classification	Tenure Months	Balance	Spend On Groceries
200000262	Zoe	Clarkson	Female	59	5	62	cotland	Other	24	23550.89	70.77
200001214	Carolyn	McDonald	Female	58	5	61.2	cotland	Other	24	69027.62	67.1
400000497	Anna	Chapman	Female	26	2	65.1	orthern Ireland	White Collar	46	5789.63	46.23
400001939	Richard	Dowd	Male	21	2	70.9	orthern Ireland	White Collar	23	10248.59	36.48
300002298	Phil	Arnold	Male	37	3	70.4	Vales	Blue Collar	15	80824.89	36.11

$$\{ 61.2, 62, 65.1, 70.4, 70.9 \}$$

$$\text{Mean} = \frac{61.2 + 62 + 65.1 + 70.4 + 70.9}{5} = 65.92$$

$$\{\ 61.2,\ 62,\ 65.1,\ 70.4,\ 70.9\ \}$$

$$\mu \quad \text{Mean} = \frac{61.2 + 62 + 65.1 + 70.4 + 70.9}{5} = 65.92$$

$$\text{Variance} = \frac{(61.2 - 65.92)^2 + (62 - 65.92)^2 + (65.1 - 65.92)^2 + (70.4 - 65.92)^2 + (70.9 - 65.92)^2}{5}$$

$$\text{Variance} = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = 16.64$$

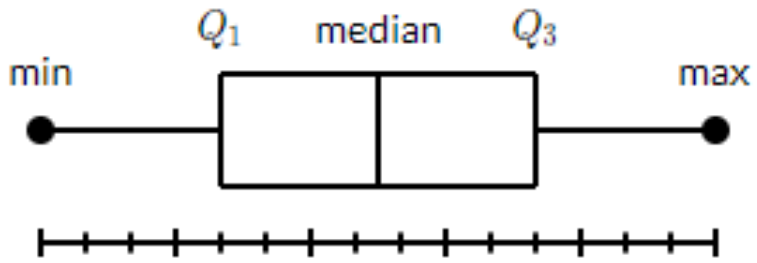
$$\text{Std. Dev.} = \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} = 4.08$$

BoxPlot

- boxplot(), shows the distribution of quantitative data in a way that facilitates comparisons between variables

vis2 = sns.boxplot(data=DF, x="IncomeGroup", y="BirthRate")

Quartiles



Q2 median

Q1 25% first quartile

Q3 third quartile

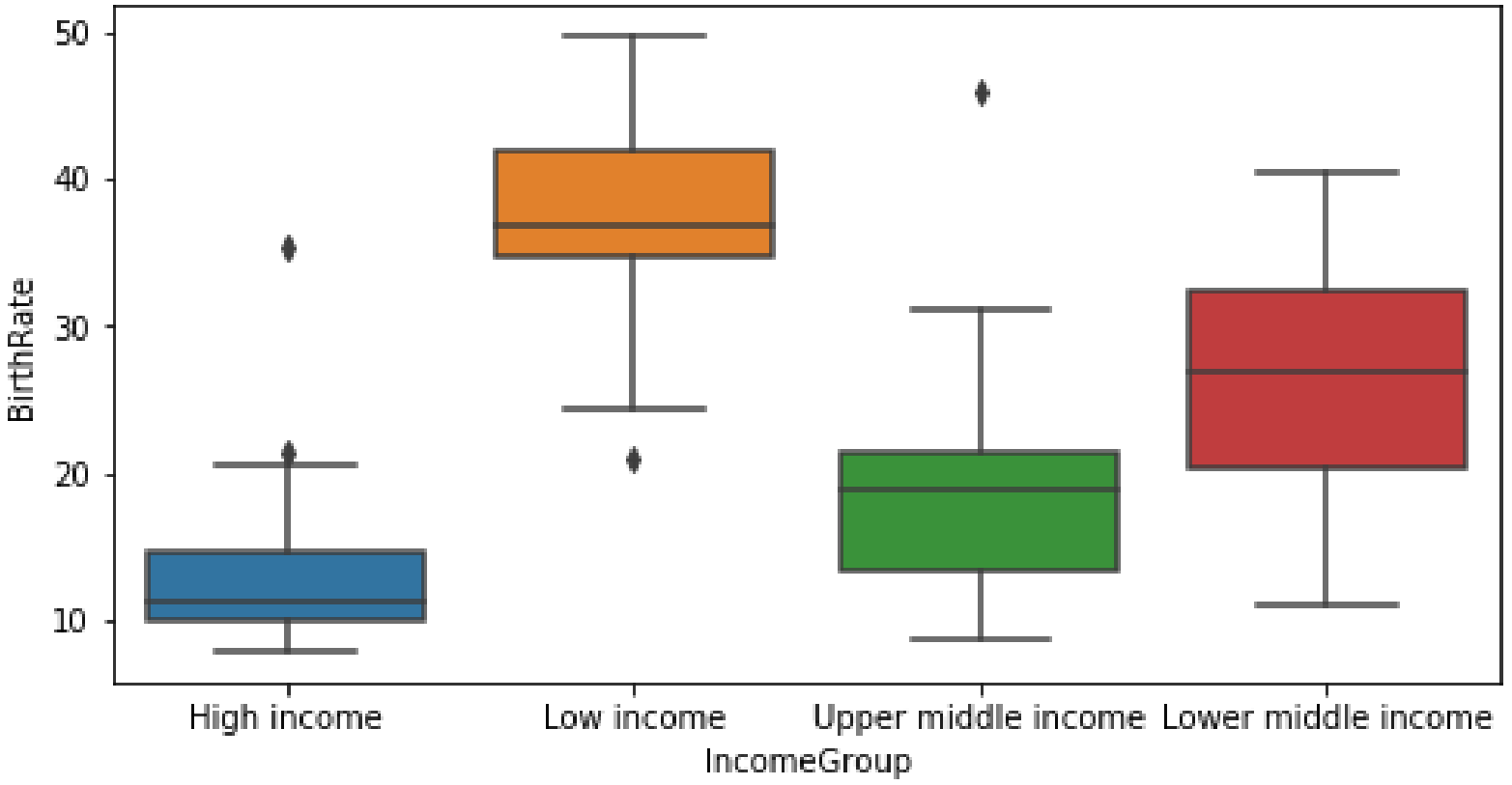
59, 60, 65, 65, 68, 69, 70, 72, 75, 75, 76, 77, 81, 82, 84, 87, 90, 95, 98

min median max

Total 19 values

The extra pointers plotted indicate outliers
(e.g. few high income rate class having high birthrate)

Task : Confirm the outliers and the plotted density values by using appropriate functions

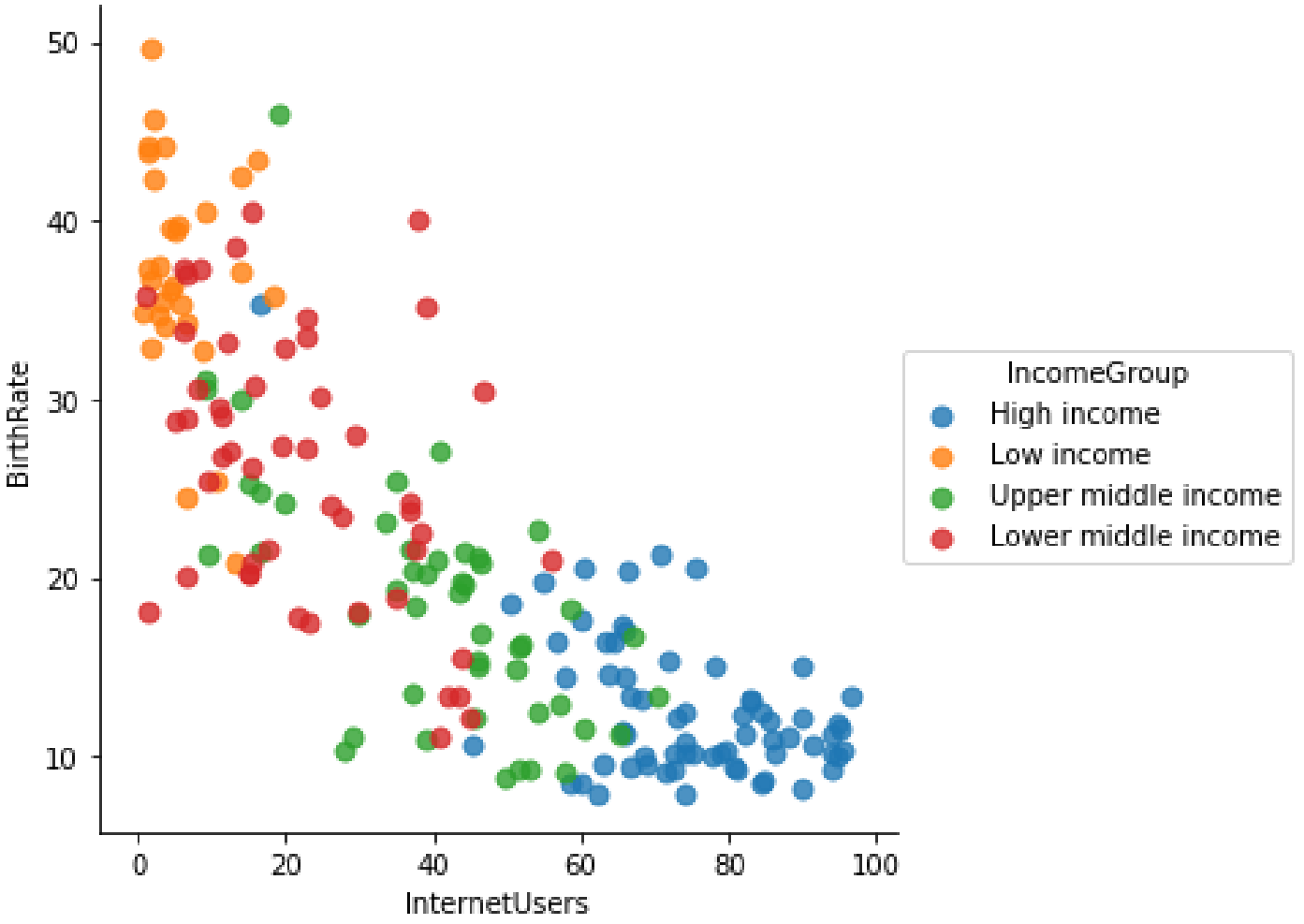


lmplot

- **BirthRate Vs Internet Users**

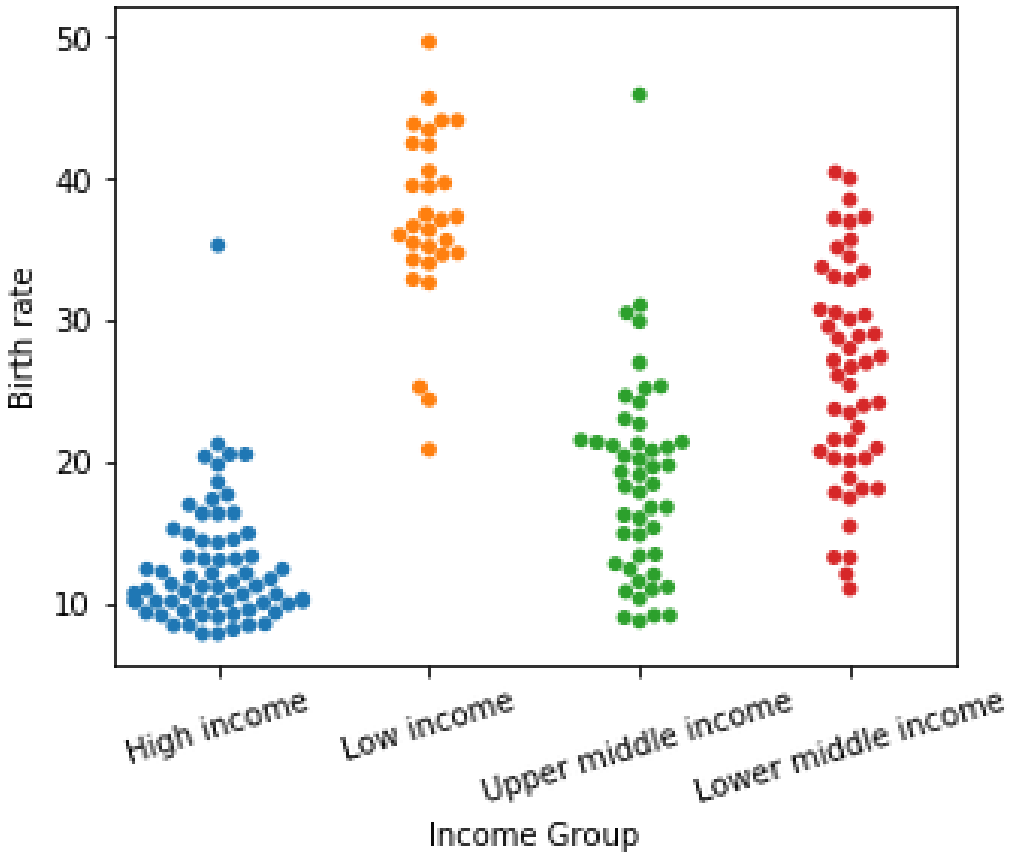
- scatter_kws is a wrapper for plt.scatter (matplotlib.pyplot.scatter), so to size the markers we need to pass value to the scatter_kws as a dictionary(key:value) , where s is the size of the marker

```
vis3 = sns.lmplot(
        data = DF, x="InternetAccess“, y="BirthRate",
        fit_reg=False,  hue="IncomeGroup", size=5)
        scatter_kws={"s":50})
```


BirthRate
50
40
30
20
10
0
20
40
60
80
100
InternetUsers
IncomeGroup
High income
Low income
Upper middle income
Lower middle income

swarmplot

```
s=sns.swarmplot(data=DF,
                x="IncomeGroup", y="BirthRate")
```



Pearson Coefficient

- **The Pearson correlation coefficient measures the linear relationship between two datasets**
- **Varies between -1 and +1 with 0 implying no correlation**
- **Correlations of -1 or +1 imply an exact linear relationship**
- **Positive correlations imply that as x increases, so does y**
- **Negative correlations imply that as x increases, y decreases**

Pearson Coefficient

$$\text{Pearsonr} = \frac{N * \text{sum}(xy) - \text{sum}(x) * \text{sum}(y)}{\text{sqrt}([N * \text{sum}(x^2) - \text{sum}(x)^2] * [N * \text{sum}(y^2) - \text{sum}(y)^2])}$$

p-Value

- The p-value roughly indicates the probability of an uncorrelated system
- The p-values are not entirely reliable but are probably reasonable for datasets larger than 500 or so
- p-value is measured with a significance level of 0.05
- p-value below 0.05 indicate correlation
- p-value above 0.05 indicate no correlation

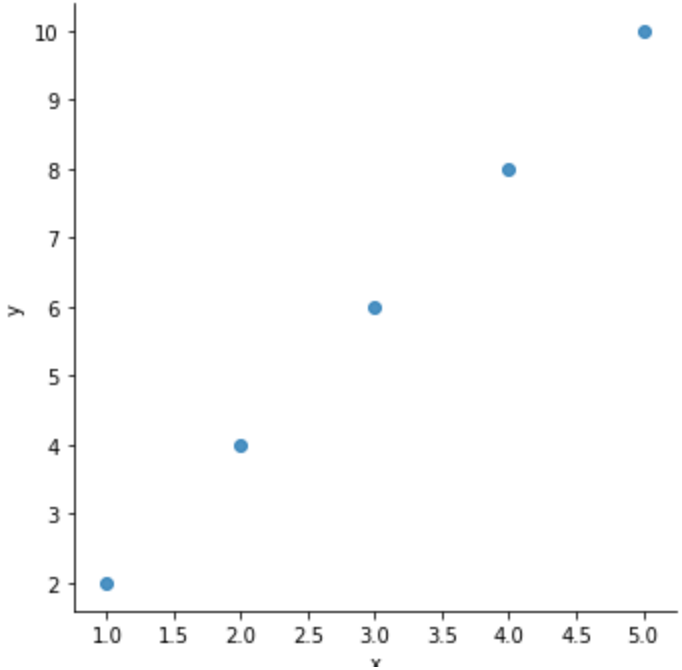
pearsonr

```
from scipy.stats.stats import pearsonr
        pearsonr([1,2,3,4,5], [2,4,6,8,10])
```

Result => (1.0)

There is a **perfect linear rel**

x & y



pearsonr

pearsonr([0,7,11,1,-5],[-2,2000,-1000,-11,0])

Result => (0.008211472)

No linear relationship

between x & y