03_Bivarijantna_analiza

August 1, 2025

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

U univarijantnoj analizi proučavamo jednu varijablu i njenu distribuciju U bivarijantnoj analizi proučavamo odnos između dvije varijable. Postoji li te ako postoji odnos,kakav je? Događa li se pri rastu jedne varijable i rast druge ili pak pad druge varijable

[2]:
```
marketing = pd.read_csv ("../../datasets/Advertising.csv")
```

[3]:
```
marketing.head()
```

[3]:

	Unnamed: 0	TV	radio	newspaper	sales
0	1	230.1	37.8	69.2	22.1
1	2	44.5	39.3	45.1	10.4
2	3	17.2	45.9	69.3	9.3
3	4	151.5	41.3	58.5	18.5
4	5	180.8	10.8	58.4	12.9

U datasetu Advertising imamo podatke o ulaganju u marketing na televiziji, radiju, novinama (jedinični iznosi) s jedne strane te posljedično prodaja (u tisućama) Želimo provjeriti događa li se rast prodaje ako ulažemo u oglašavanje

[4]:
```
marketing.describe()
```

[4]:

	Unnamed: 0	TV	radio	newspaper	sales
count	200.000000	200.000000	200.000000	200.000000	200.000000
mean	100.500000	147.042500	23.264000	30.554000	14.022500
std	57.879185	85.854236	14.846809	21.778621	5.217457
min	1.000000	0.700000	0.000000	0.300000	1.600000
25%	50.750000	74.375000	9.975000	12.750000	10.375000
50%	100.500000	149.750000	22.900000	25.750000	12.900000
75%	150.250000	218.825000	36.525000	45.100000	17.400000
max	200.000000	296.400000	49.600000	114.000000	27.000000

[5]:
```
marketing.corr(numeric_only=True)
```

[5]:

	Unnamed: 0	TV	radio	newspaper	sales
Unnamed: 0	1.000000	0.017715	-0.110680	-0.154944	-0.051616

TV	0.017715	1.000000	0.054809	0.056648	0.782224
radio	-0.110680	0.054809	1.000000	0.354104	0.576223
newspaper	-0.154944	0.056648	0.354104	1.000000	0.228299
sales	-0.051616	0.782224	0.576223	0.228299	1.000000

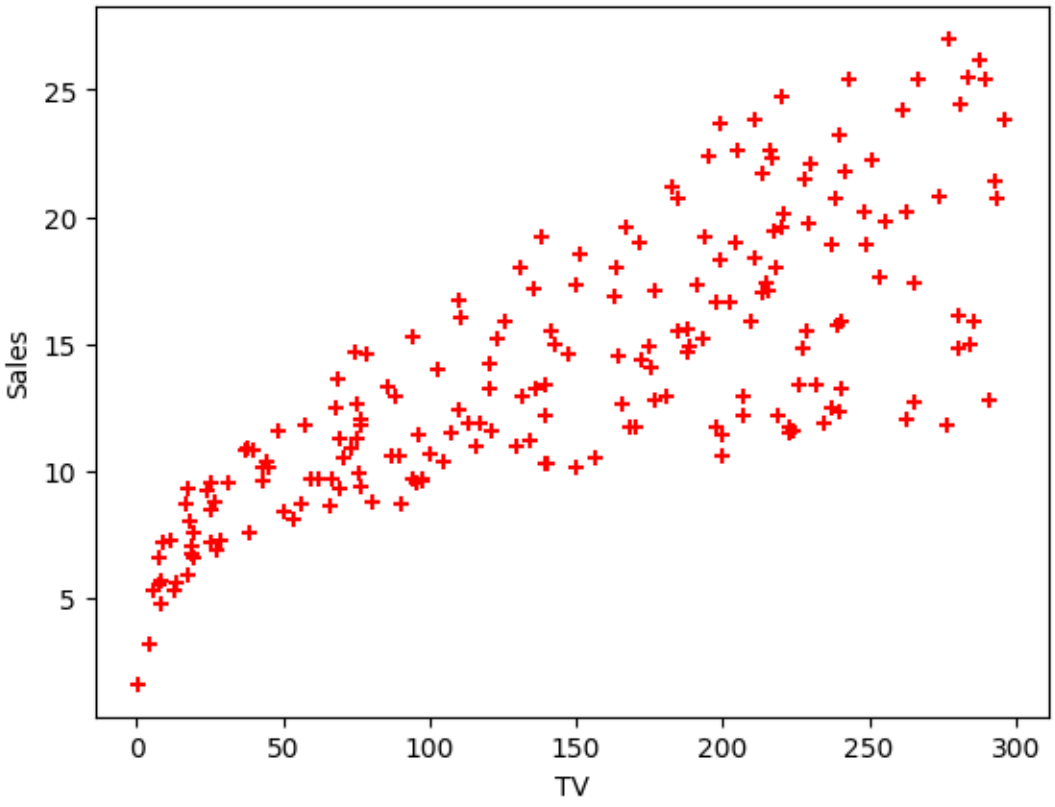
Statistička mjera koja nam govori postoji li veza među varijablama zove se korelacija

1 Zadatak 1: u kojem rasponu se kreće korelacija?

Dajte po jedan primjer pozitivne korelacije i negativne korelacije! Analizirajte prošlu tablicu, kolika je povezanost između oglašavanja na televiziji i prodaje, oglašavanja u novinama i prodaje te oglašavanja na radiju i prodaje?

[6]:
```
plt.xlabel('TV')
plt.ylabel('Sales')
plt.scatter(marketing.TV, marketing.sales, color='red', marker='+')
```

[6]: <matplotlib.collections.PathCollection at 0x7f31b5d52c10>



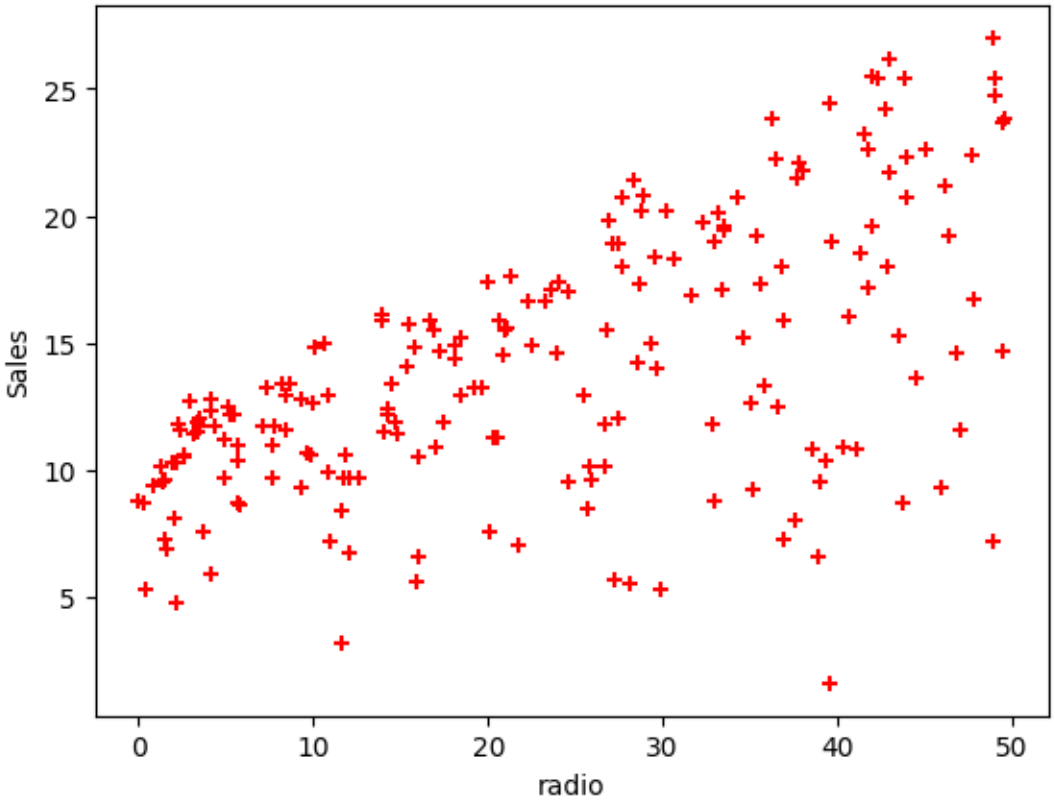
Veza među varijablama se može i grafički prikazati korištenjem scatter dijagrama. Za scatter dijagram trebamo dvije varijable Objasnite grafikon, što vidimo na njemu

2 Zadatak 2

napravite scatter dijagrame u kome ćete prikazati odnos između novina i prodaje te radija i prodaje

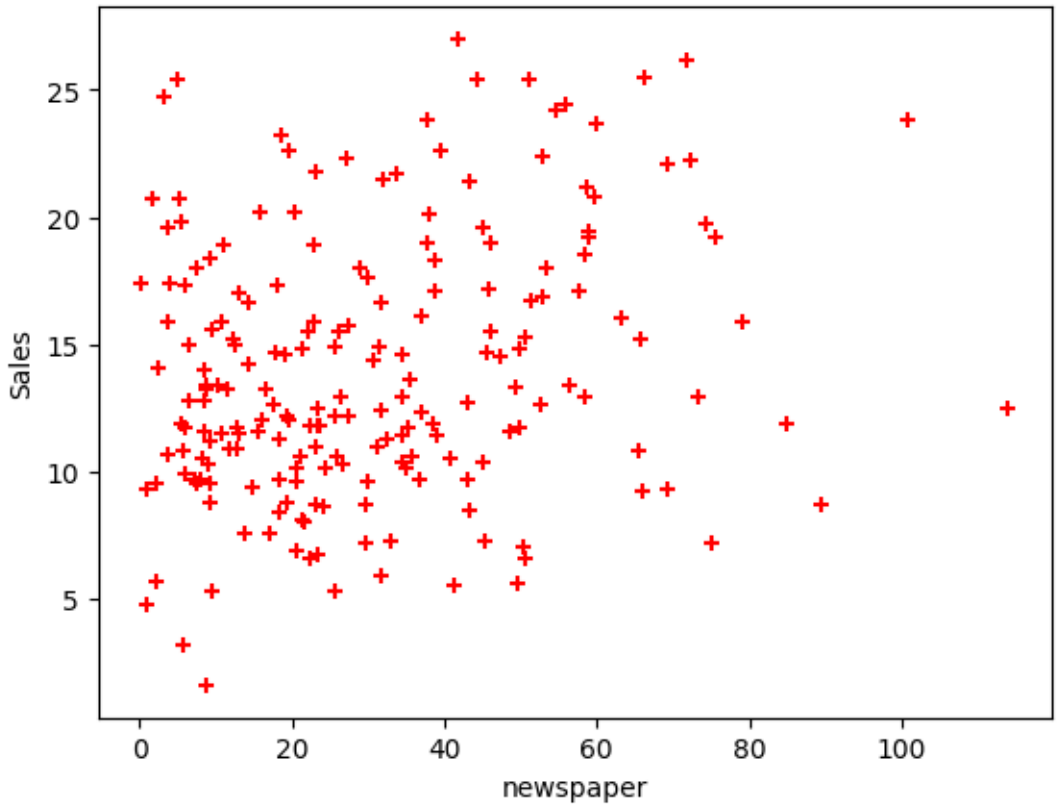
[7]:
```
plt.xlabel('radio')
plt.ylabel('Sales')
plt.scatter(marketing.radio, marketing.sales, color='red', marker='+')
```

[7]: <matplotlib.collections.PathCollection at 0x7f31b3b14bd0>



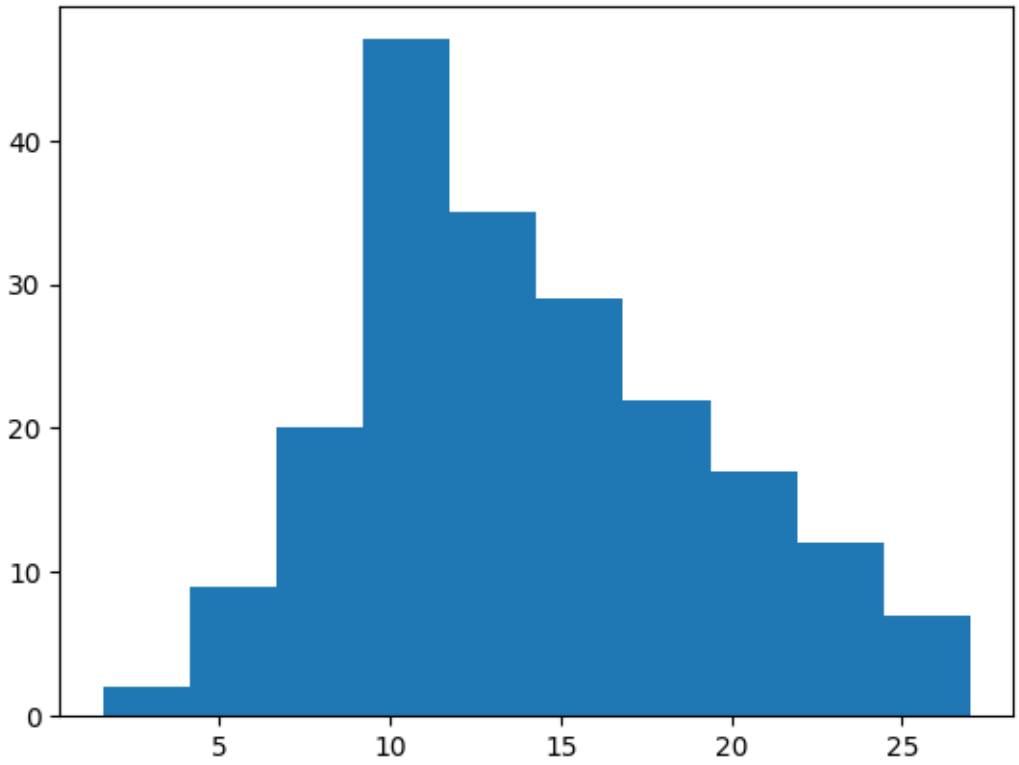
[8]:
```
plt.xlabel('newspaper')
plt.ylabel('Sales')
plt.scatter(marketing.newspaper, marketing.sales, color='red', marker='+')
```

[8]: <matplotlib.collections.PathCollection at 0x7f31b5de3dd0>



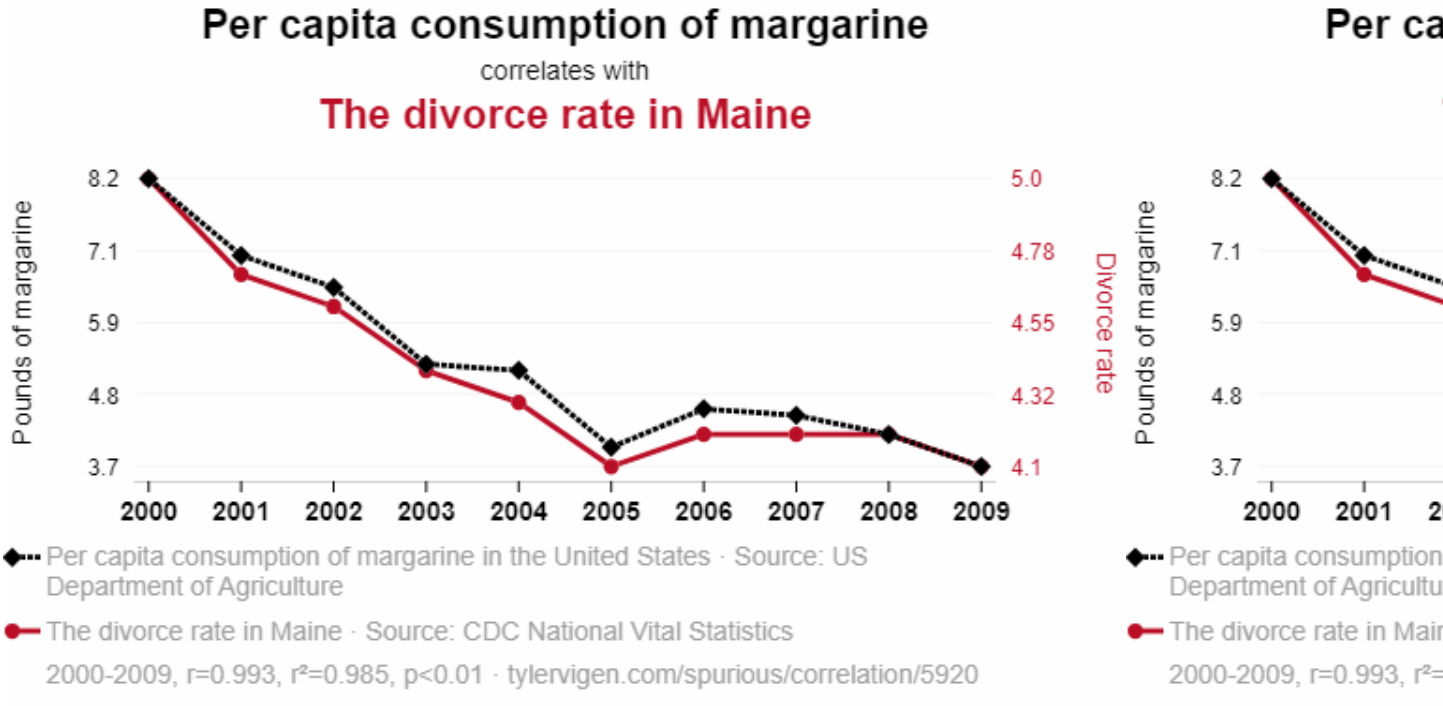
[9]:
```
plt.hist (marketing.sales)
```

[9]:
```
(array([ 2.,  9., 20., 47., 35., 29., 22., 17., 12.,  7.]),
 array([ 1.6 ,  4.14,  6.68,  9.22, 11.76, 14.3 , 16.84, 19.38, 21.92,
        24.46, 27.  ]),
 <BarContainer object of 10 artists>)
```



Ako otkrijemo korelaciju između dvije varijable, možemo li baš svaki put reći da postoji veza? Moramo li koji put uključiti i domain knowledge expertise?

3 Primjeri korelacija koji nemaju smisla



Još primjera se može naći na ovom linku Spurious correlations :(https://www.tylervigen.com/spurious-correlations)

4 Zadatak 3

učitajte notebook u kome obrađujete Titanic dataset. napravite scatter dijagrame za varijable Age & Fare te Pclass & Fare. Postoje li veze među tim varijablama?

[10]:
```
marketing.corr(numeric_only = True)
```

[10]:

	Unnamed: 0	TV	radio	newspaper	sales
Unnamed: 0	1.000000	0.017715	-0.110680	-0.154944	-0.051616
TV	0.017715	1.000000	0.054809	0.056648	0.782224
radio	-0.110680	0.054809	1.000000	0.354104	0.576223
newspaper	-0.154944	0.056648	0.354104	1.000000	0.228299
sales	-0.051616	0.782224	0.576223	0.228299	1.000000

Upozorenje: ponekad pri provođenju EDA, dolazi do pogreške uzorkovane lijenošću. Koristi se samo naredba corr() za otkrivanje korelacije, bez korištenja grafičkog prikaza za bivarijantnu analizu. Može doći do ozbiljne pogreške jer nam naredba corr() ne pokazuje kako su podaci distribuirani.

[11]:
```
df1 = pd.read_excel ("../../datasets/Anscombe.xlsx", sheet_name = "prvi")
df2 = pd.read_excel ("../../datasets/Anscombe.xlsx", sheet_name = "drugi")
df3 = pd.read_excel ("../../datasets/Anscombe.xlsx", sheet_name = "treci")
df4 = pd.read_excel ("../../datasets/Anscombe.xlsx", sheet_name = "cetvrti")
```

Učitavamo Excel file Anscombe koji ima četiri radna lista, na svakom je posebna tablica. Svaka tablica se učitava u zaseban dataset df1, df2, df3 i df4.

[12]:
```
df1.head()
```

[12]:

	x	y
0	10	8.04
1	8	6.95
2	13	7.58
3	9	8.81
4	11	8.33

[13]:
```
df2.head()
```

[13]:

	x	y
0	10	9.14
1	8	8.14
2	13	8.74
3	9	8.77
4	11	9.26

[14]:
```
df1.describe()
```

[14]:

	x	y
count	11.000000	11.000000
mean	9.000000	7.500909
std	3.316625	2.031568
min	4.000000	4.260000
25%	6.500000	6.315000
50%	9.000000	7.580000
75%	11.500000	8.570000
max	14.000000	10.840000

[15]:
```
df2.describe()
```

[15]:

	x	y
count	11.000000	11.000000
mean	9.000000	7.500909
std	3.316625	2.031657
min	4.000000	3.100000
25%	6.500000	6.695000
50%	9.000000	8.140000
75%	11.500000	8.950000
max	14.000000	9.260000

[16]:
```
df3.describe()
```

[16]:

	x	y
count	11.000000	11.000000
mean	9.000000	7.500000
std	3.316625	2.030424
min	4.000000	5.390000
25%	6.500000	6.250000
50%	9.000000	7.110000
75%	11.500000	7.980000
max	14.000000	12.740000

[17]:
```
df4.describe()
```

[17]:

	x	y
count	11.000000	11.000000
mean	9.000000	7.500909
std	3.316625	2.030579
min	8.000000	5.250000
25%	8.000000	6.170000
50%	8.000000	7.040000
75%	8.000000	8.190000
max	19.000000	12.500000

Sva četiri dataseta imaju isti prosjek i standardnu devijaciju

[18]:
```
df1.corr()
```

[18]:

	x	y
x	1.000000	0.816421
y	0.816421	1.000000

[19]:
```
df2.corr()
```

[19]:

	x	y
x	1.000000	0.816237
y	0.816237	1.000000

[20]:
```
df3.corr()
```

[20]:

	x	y
x	1.000000	0.816287
y	0.816287	1.000000

[21]:
```
df4.corr()
```

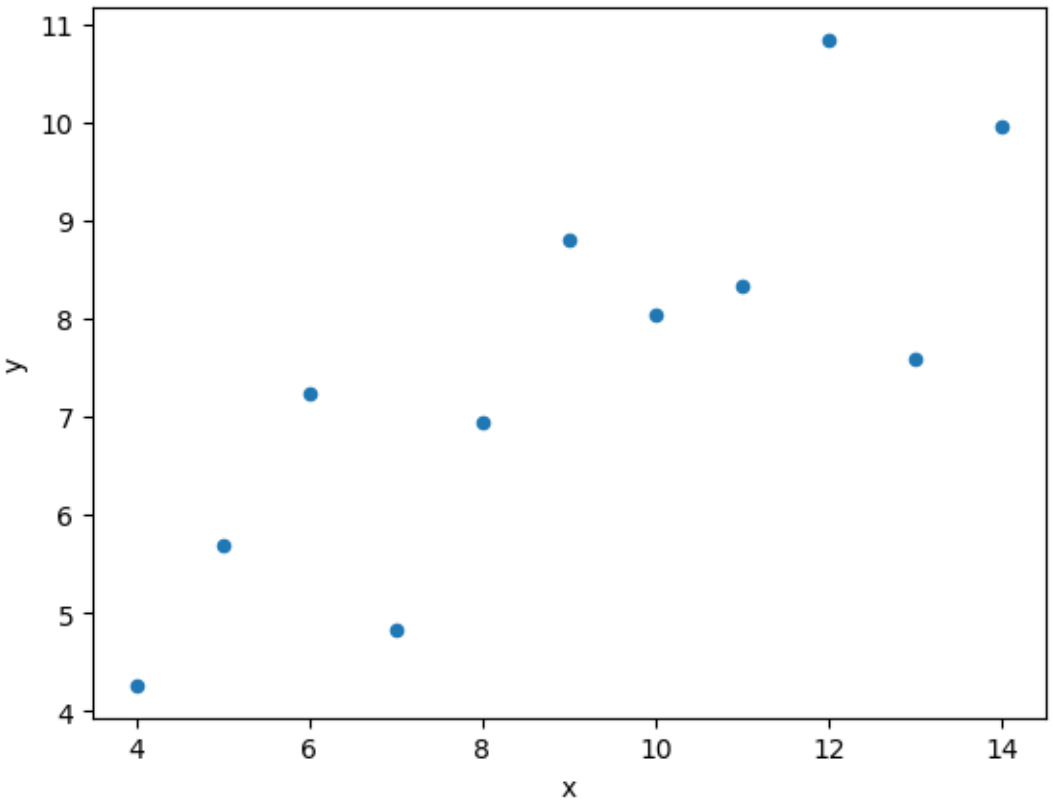
[21]:

	x	y
x	1.000000	0.816521
y	0.816521	1.000000

Sva četiri dataseta imaju i isti koeficijent korelacije. Super, možemo zaključiti da postoji veza između varijable x i y, savršeno za predikciju (linearna regresia rulles!) Međutim, ako grafički usporedimo varijable x i y u sva četiri dataseta...

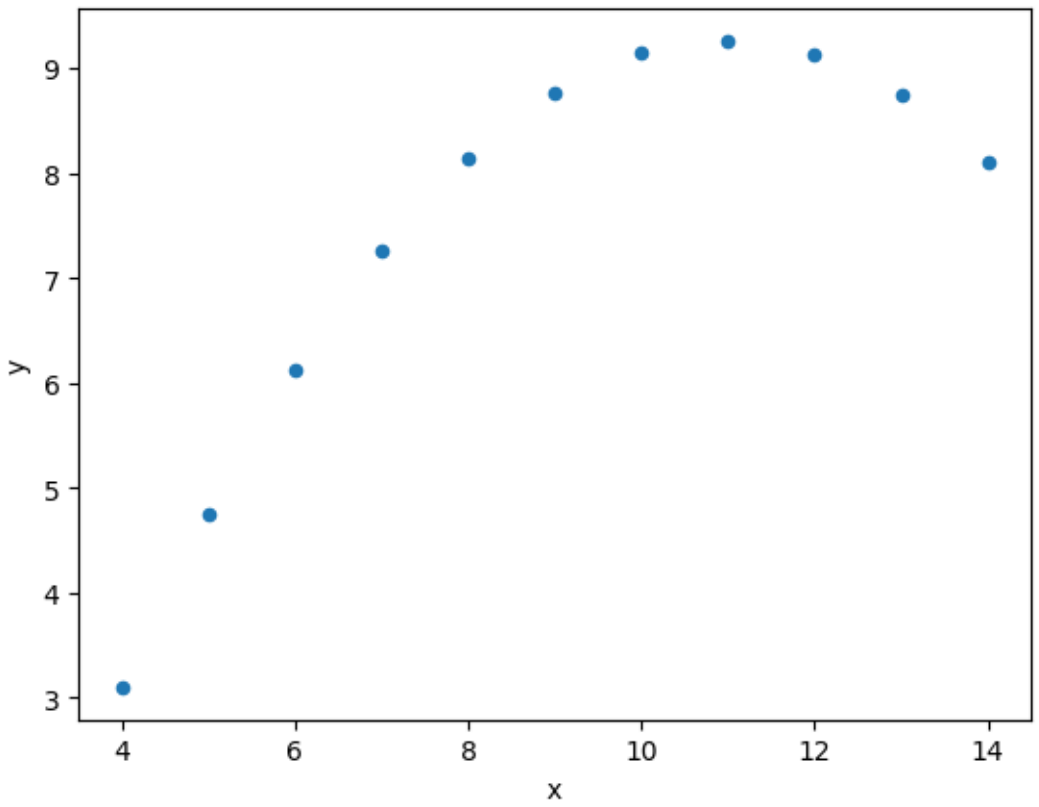
[22]:
```
df1.plot (kind = "scatter", x = "x", y = "y")
```

[22]:
```
<Axes: xlabel='x', ylabel='y'>
```



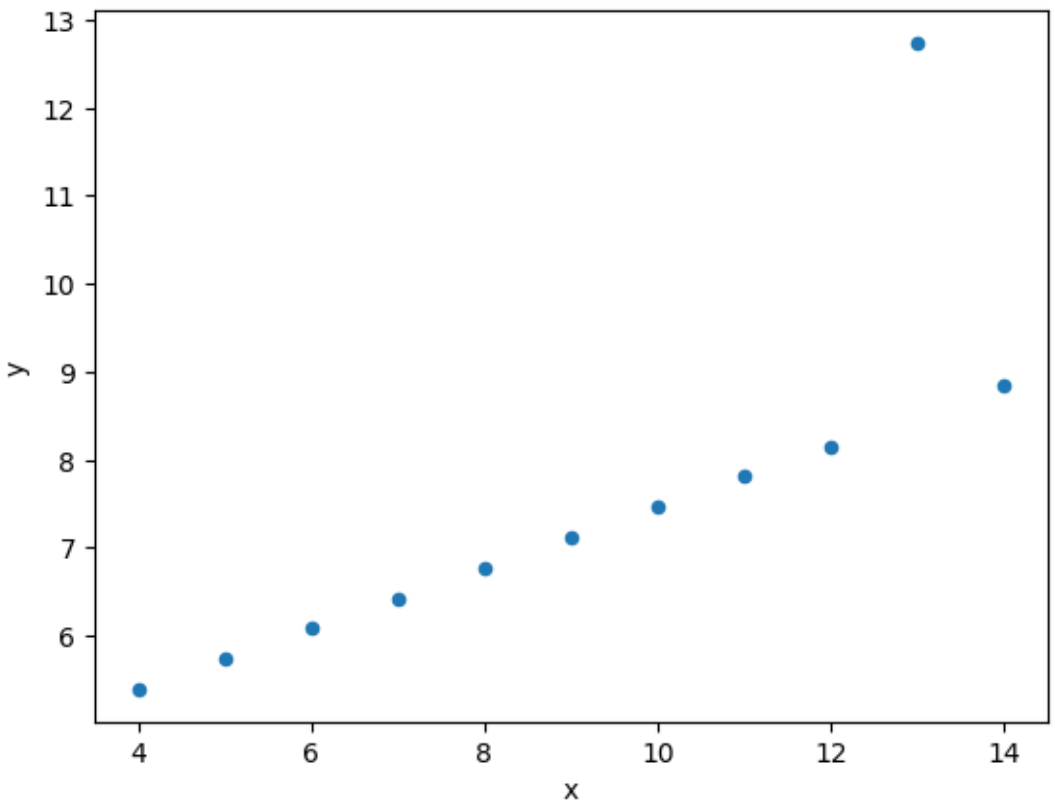
[23]:
```
df2.plot (kind = "scatter", x = "x", y = "y")
```

[23]:
```
<Axes: xlabel='x', ylabel='y'>
```



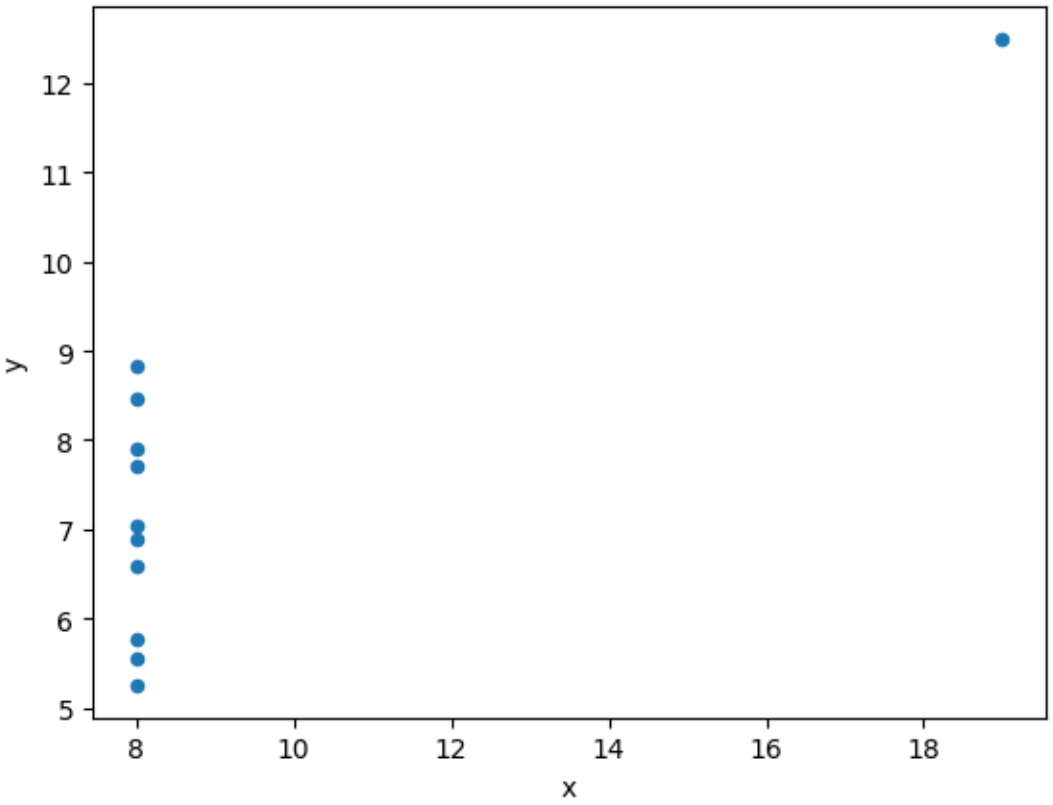
[24]:
```
df3.plot (kind = "scatter", x = "x", y = "y")
```

[24]: <Axes: xlabel='x', ylabel='y'>



[25]:
```
df4.plot (kind = "scatter", x = "x", y = "y")
```

[25]: <Axes: xlabel='x', ylabel='y'>



Ova 4 dataseta (popularno zvana Anscombeov kvartet) nam pokazuju važnost vizualizacije podataka kako se ne bi samo koristile statističke mjere

5 correlation does not mean causation!