HW-Topic-7

Data Acquisition, Modeling and Analysis: Big Data Analytics

Submitted By – **Sudhanshu Kakkar**
CWID – **20036779**

Sudhanshu Kakkar | ID - 20036779

Correlation

WHAT IS IT?

- A statistical measure that describes the relationship between two or more variables.
- It lets us understand the relatedness between two variables, allowing for testing against models or simplification of analysis

TYPES

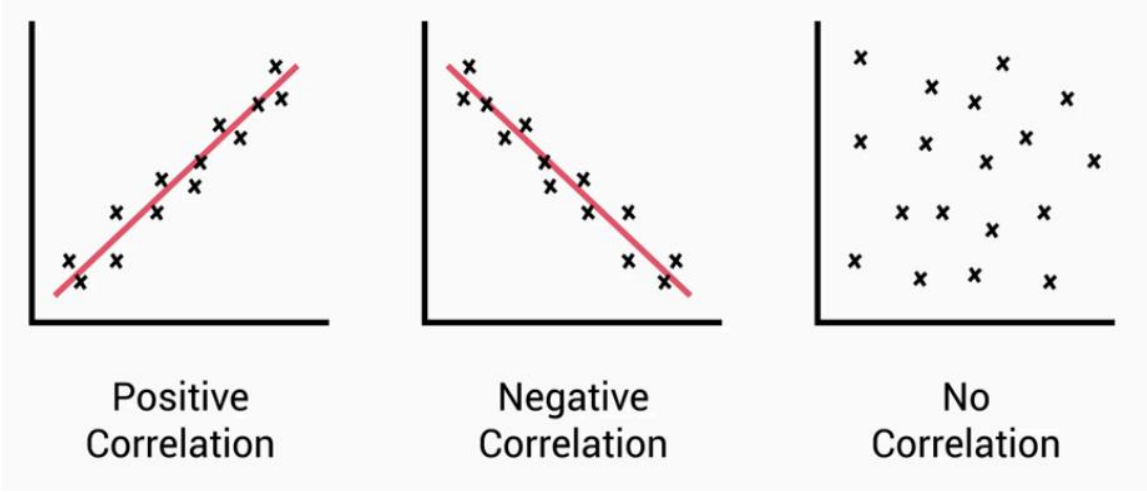
- **Positive correlation**:
  When one variable increases, so does the other.
- **Negative correlation**:
  When one variable increases, the other variable decreases.
- **No correlation**:
  When there is no linear relationship between variables.

APPLICATIONS

- Feature Selection in Data Science
- Dimensionality Reduction in case of correlated features

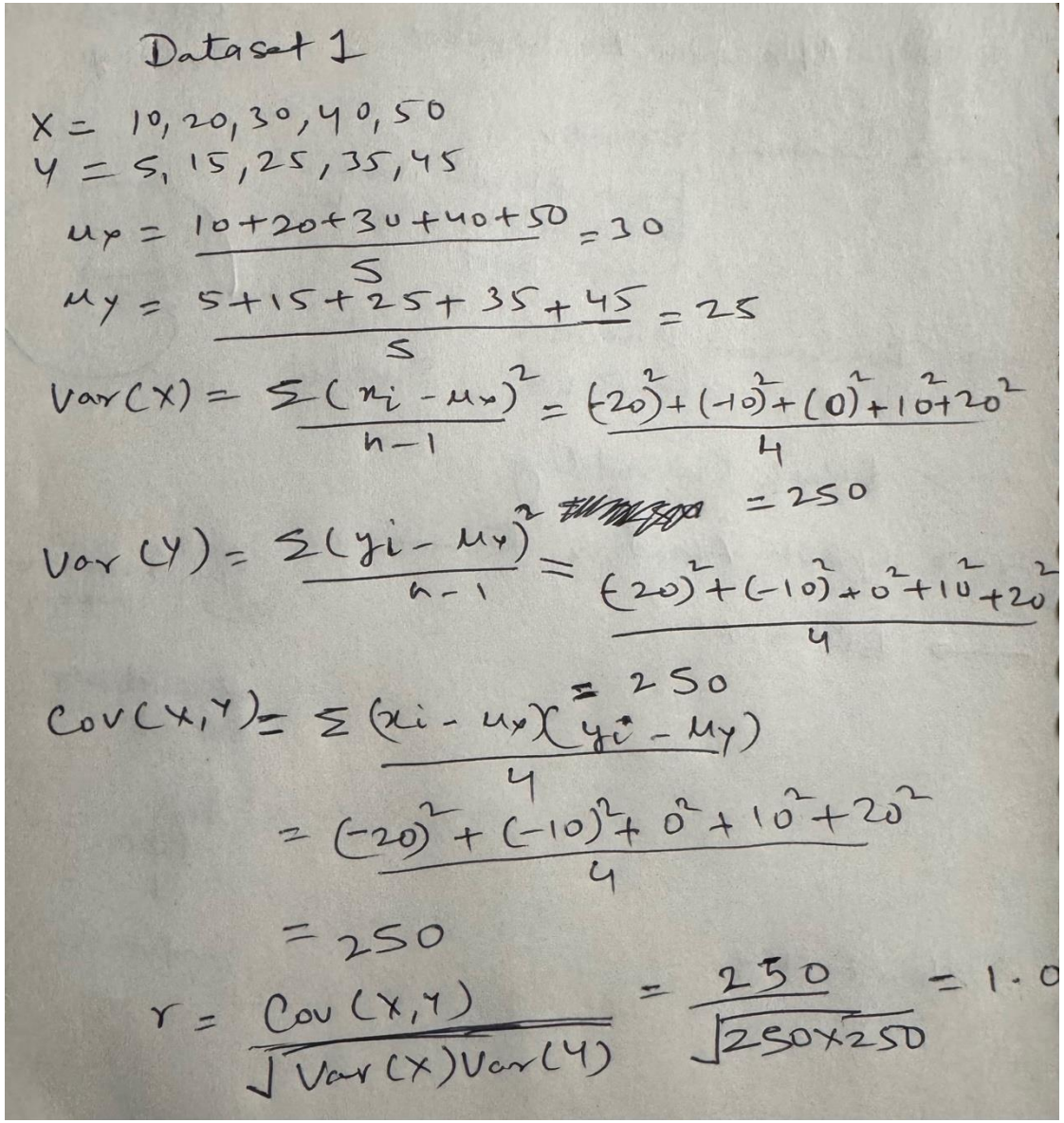
The Correlation Scale

- Range of -1 to 1
- 1 = strong positive correlation
- 0 to 1 = some positive correlation
- 0 = no correlation
- -1 to 0 = some negative correlation
- -1 = strong negative correlation



Manual calculation of Correlation Coefficient

Sudhanshu Kakkar | ID - 20036779



Dataset 1

$X = 10, 20, 30, 40, 50$

$Y = 5, 15, 25, 35, 45$

$\mu_x = \frac{10+20+30+40+50}{5} = 30$

$\mu_y = \frac{5+15+25+35+45}{5} = 25$

$Var(X) = \frac{\Sigma(x_i - \mu_x)^2}{n-1} = \frac{(-20)^2+(-10)^2+(0)^2+10^2+20^2}{4} = 250$

$Var(Y) = \frac{\Sigma(y_i - \mu_y)^2}{n-1} = \frac{(-20)^2+(-10)^2+0^2+10^2+20^2}{4} = 250$

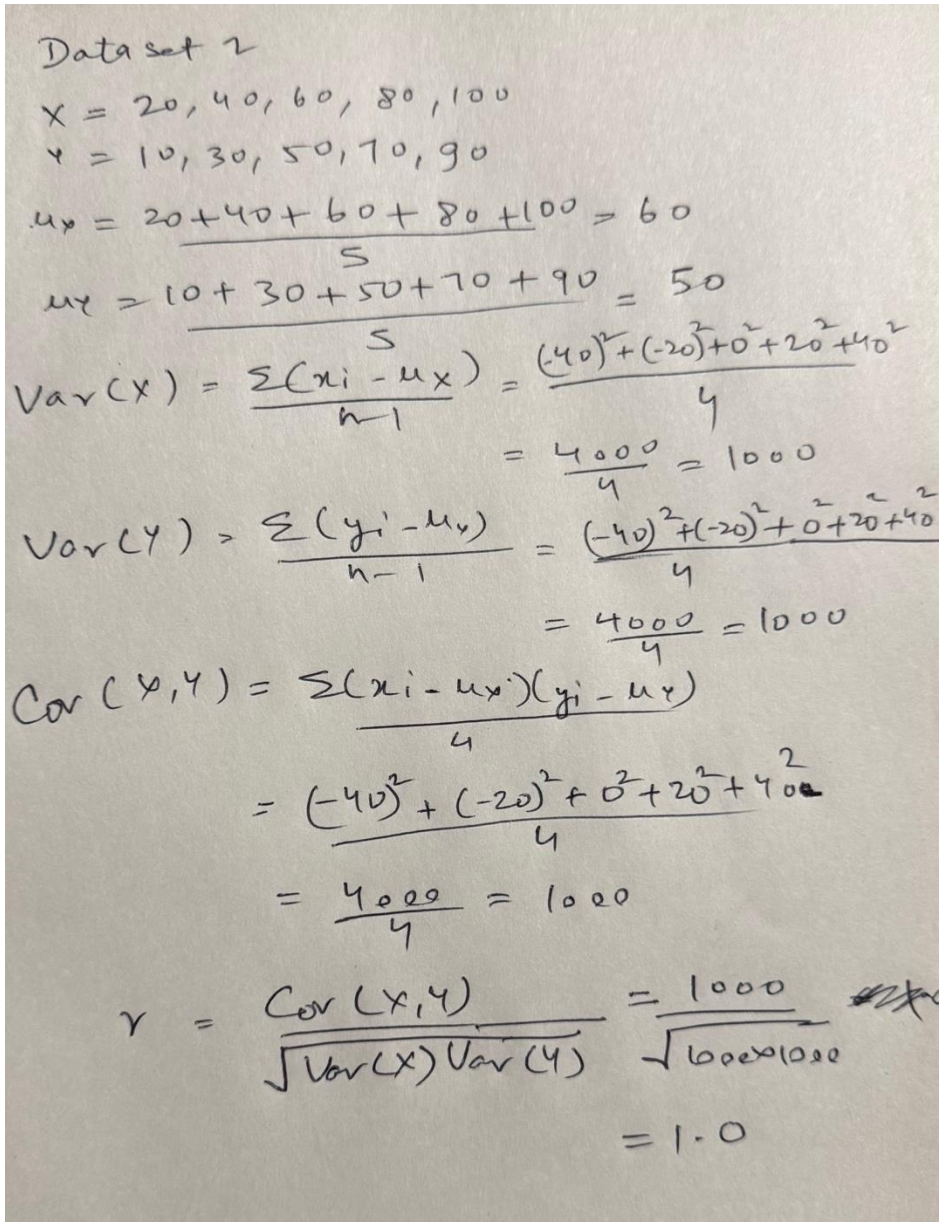
$Cov(X,Y) = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{4}$

$= \frac{(-20)^2 + (-10)^2 + 0^2 + 10^2 + 20^2}{4}$

$= 250$

$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{250}{\sqrt{250 \times 250}} = 1.0$

Solution for Dataset 1



Dataset 2

$X = 20, 40, 60, 80, 100$

$Y = 10, 30, 50, 70, 90$

$\mu_x = \frac{20+40+60+80+100}{5} = 60$

$\mu_y = \frac{10+30+50+70+90}{5} = 50$

$Var(X) = \frac{\Sigma(x_i - \mu_x)}{n-1} = \frac{(-40)^2+(-20)^2+0^2+20^2+40^2}{4}$

$= \frac{4000}{4} = 1000$

$Var(Y) = \frac{\Sigma(y_i - \mu_y)}{n-1} = \frac{(-40)^2+(-20)^2+0^2+20^2+40^2}{4}$

$= \frac{4000}{4} = 1000$

$Cov(X,Y) = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{4}$

$= \frac{(-40)^2 + (-20)^2 + 0^2 + 20^2 + 40^2}{4}$

$= \frac{4000}{4} = 1000$

$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{1000}{\sqrt{1000 \times 1000}}$

$= 1.0$

Solution for Dataset 2

Sudhanshu Kakkar | ID - 20036779

```python
import numpy as np

def CorrCoef(X,Y) :
    #  calculating mean of X and Y
    mux = np.mean (X)
    muy = np. mean (Y)

    # calculating variance for X and Y
    VarX = np. sum ( (X-mux)**2)/(len(X)-1)
    VarY = np. sum ( (Y-muy)**2)/(len(Y)-1)

    # calculating covariance of X and Y
    CoV = np. sum ( (X-mux)*(Y-muy))/(len(X)-1)

    # calculating correlating coeff here
    CoCo = CoV/np.sqrt(VarX*VarY)

    return CoCo

# Dataset 1
X = [10,20,30,40,50]
Y = [5,15,25,35,45]

C = CorrCoef(X,Y)
print(f"The Correlation Coefficient for dataset 1 is {C:.2f}")

# Dataset 2
x = [20,40,60,80,100]
Y = [10,30,50,70,90]
C = CorrCoef(X,Y)

print(f"The Correlation Coefficient for dataset 2 is {C:.2f}")
```

```
The Correlation Coefficient for dataset 1 is 1.00
The Correlation Coefficient for dataset 2 is 1.00
```

Program to calculate: Correlation Coefficient