STATISTICS REVIEW

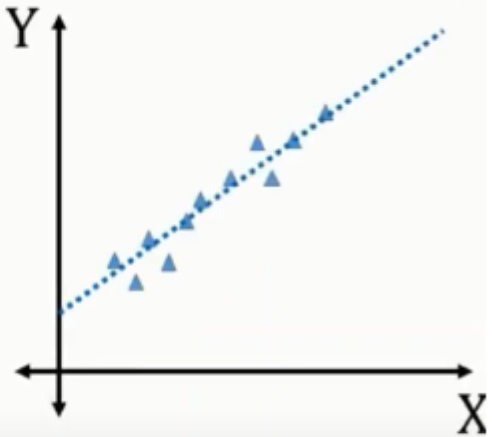
- **Correlation**
- Variance , Covariance and correlation

- **Correlation**
- Variance , Covariance and correlation

Correlation

Statistically Correlated

- Strength of the correlation – Coefficient of Correlation
- Direction of correlation – Sign of the Coefficient



Pearson Correlation Coefficient

$$r = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{(N - 1) * \sigma_x * \sigma_y}$$

X=Height

Y=Weight

Correlation Coefficient

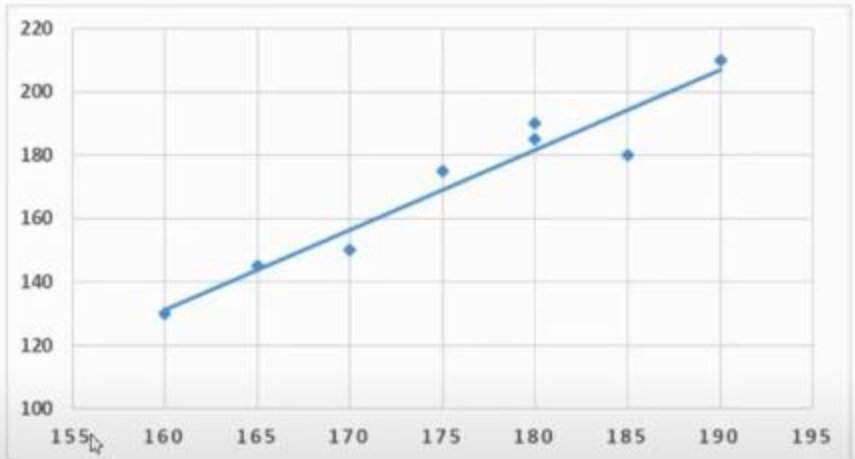
	Height X	Weight Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})*(Y-\bar{Y})$
	160	130	-15.625	-40.625	634.7656
	170	150	-5.625	-20.625	116.0156
	165	145	-10.625	-25.625	272.2656
	180	190	4.375	19.375	84.76563
	175	175	-0.625	4.375	-2.73438
	190	210	14.375	39.375	566.0156
	185	180	9.375	9.375	87.89063
	180	185	4.375	14.375	62.89063
Mean	**175.625**	**170.625**			**1821.875**
Std Dev	**10.155**	**25.651**			

$$r = \frac{\sum (x-\bar{x})*(y-\bar{y})}{(N-1)*\sigma_x*\sigma_y}$$

$$r = \frac{1821.875}{(8-1)*10.155*25.651}$$

$$r = 0.96$$

Correlation Coefficient

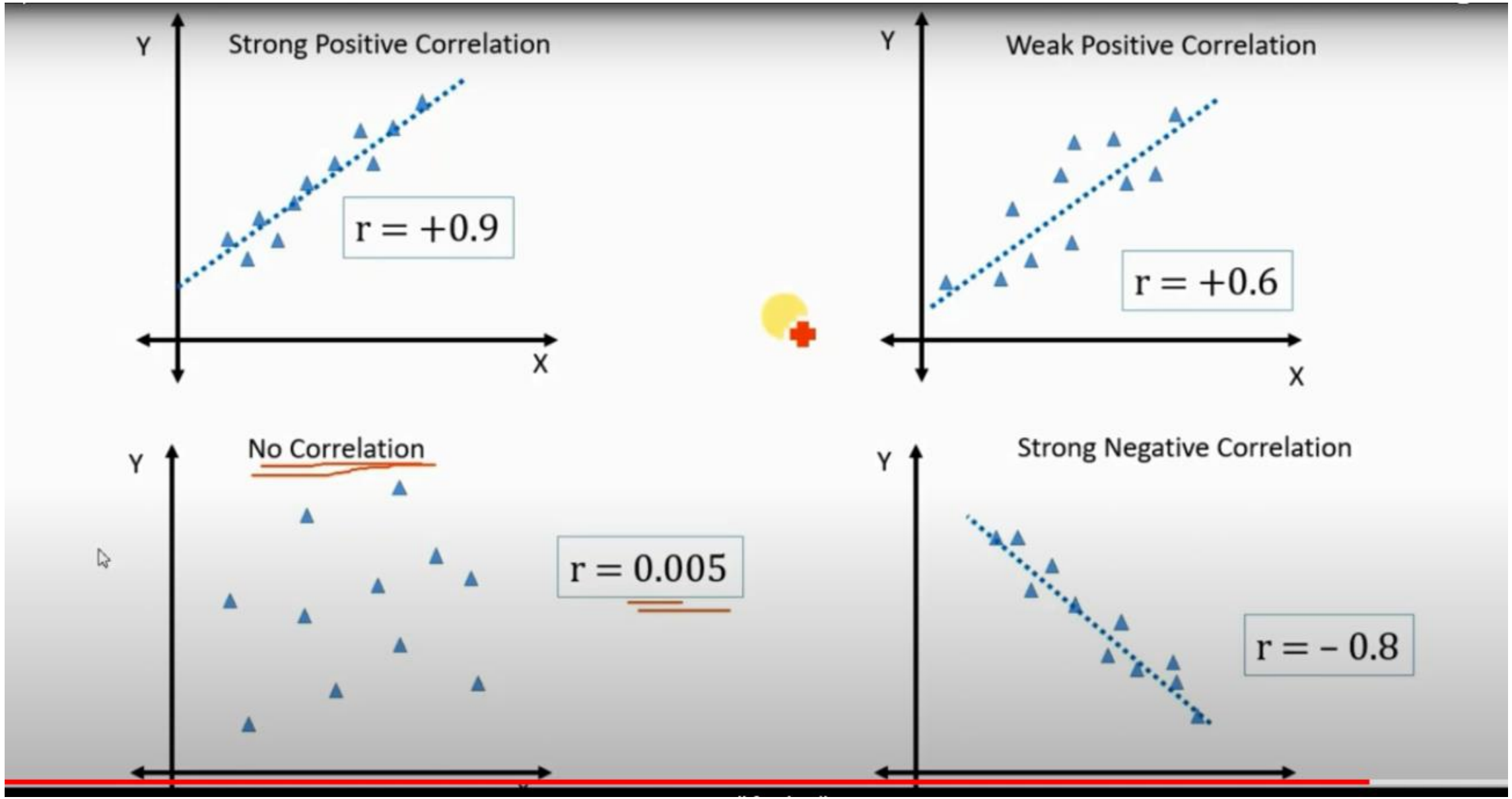


Scatter Plot

$$r = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{(N - 1) * \sigma_x * \sigma_y}$$

$$r = \frac{1821.875}{(8-1) * 10.155 * 25.651}$$

$$r = 0.96$$

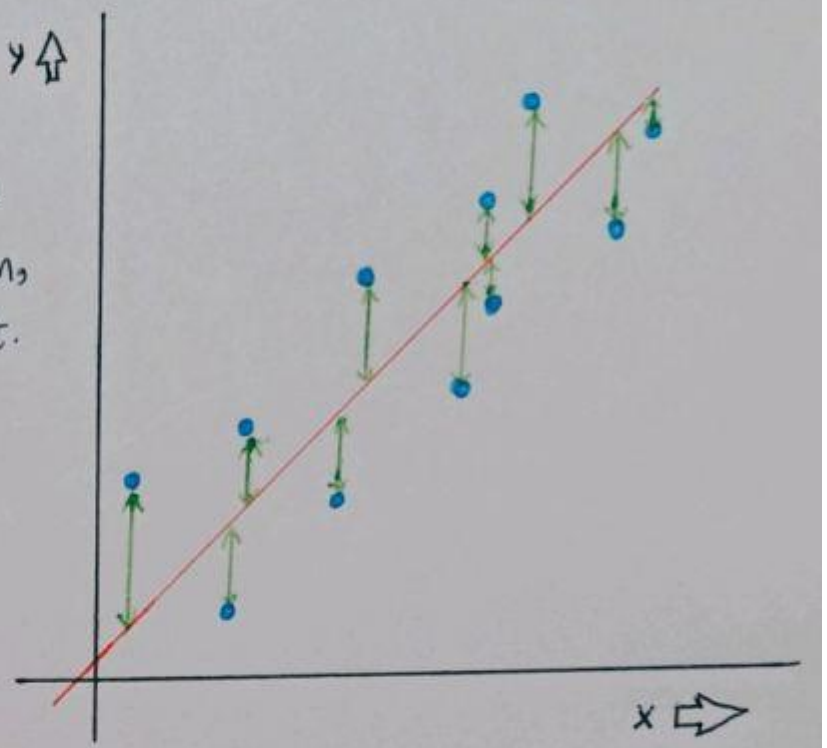

Y
Strong Positive Correlation
r = +0.9
X
Y
Weak Positive Correlation
r = +0.6
X
Y
No Correlation
r = 0.005
Y
Strong Negative Correlation
r = – 0.8

Variance & Standard deviation

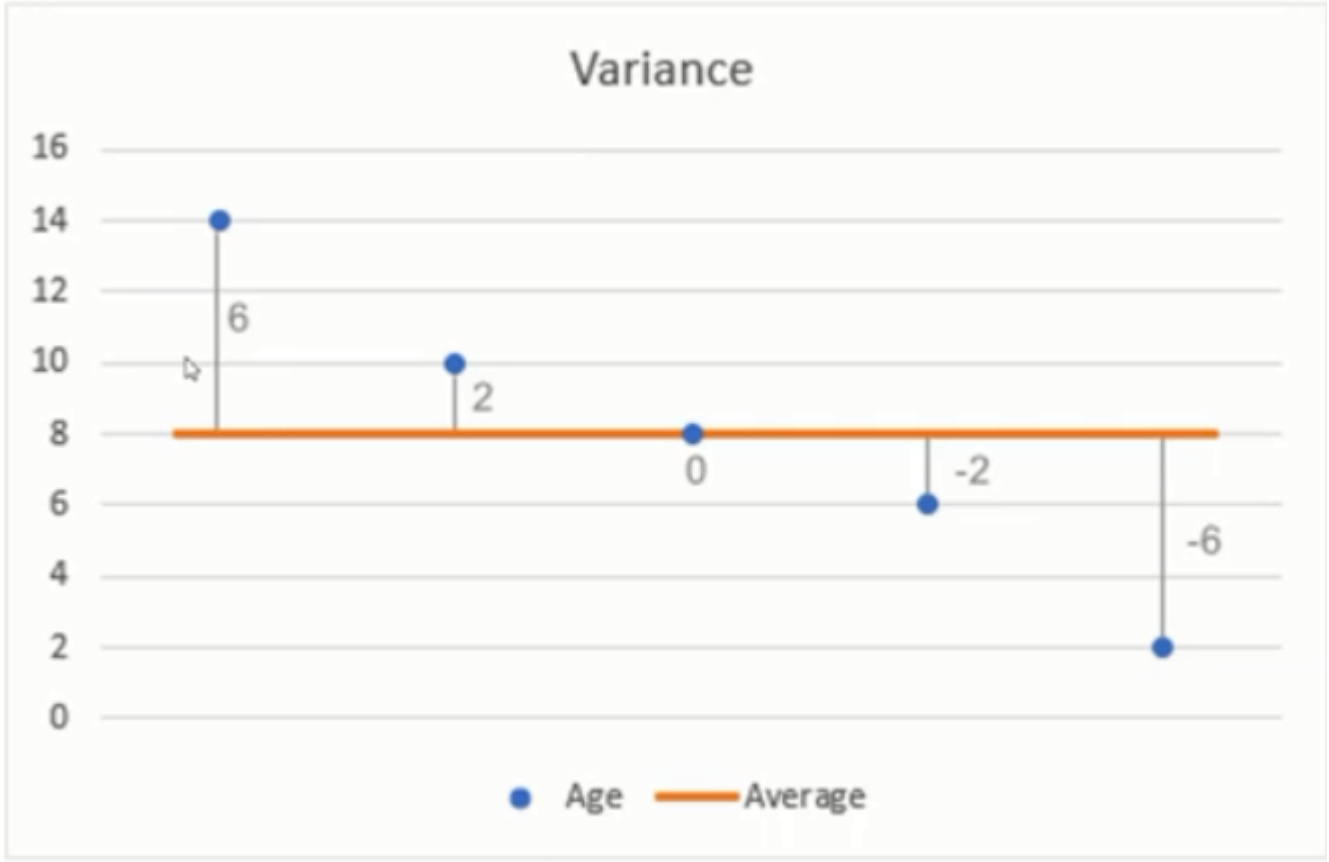
Variance measures how far each number in the set is from the mean, and thus every other number in the set.

$$\text{Sample Variance } \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N-1}$$

$$\text{Standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N-1}}$$



Variance , Covariance and correlation



$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

s^2 = sample variance

x_i = value of i^{th} element

$\bar{x}$ = sample mean

n = sample size

How To Calculate Variance

Data	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
5	−4	16
6	−3	9
8	−1	1
9	0	0
10	1	1
11	2	4
14	5	25

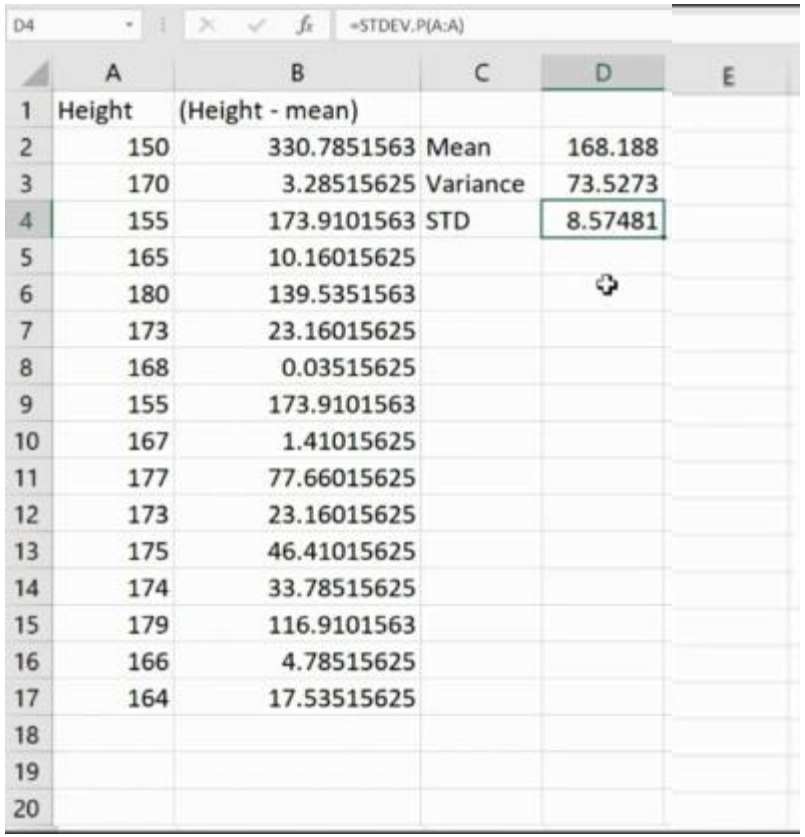
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Variance & St. Deviation

X	$\bar{X}$	$X-\bar{X}$	$(X-\bar{X})^2$
6	8	-2	4
7	8	-1	1
8	8	0	0
9	8	+1	1
10	8	+2	4

$$\bar{X} = \text{Sum} / n$$

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$



	A	B	C	D	E
1	Height	(Height - mean)			
2	150	330.7851563	Mean	168.188	
3	170	3.28515625	Variance	73.5273	
4	155	173.9101563	STD	8.57481	
5	165	10.16015625			
6	180	139.5351563			
7	173	23.16015625			
8	168	0.03515625			
9	155	173.9101563			
10	167	1.41015625			
11	177	77.66015625			
12	173	23.16015625			
13	175	46.41015625			
14	174	33.78515625			
15	179	116.9101563			
16	166	4.78515625			
17	164	17.53515625			
18					
19					
20					

Covariance

Covariance is a measure of the relation between two random variables: x and y and to what extent, they change together.

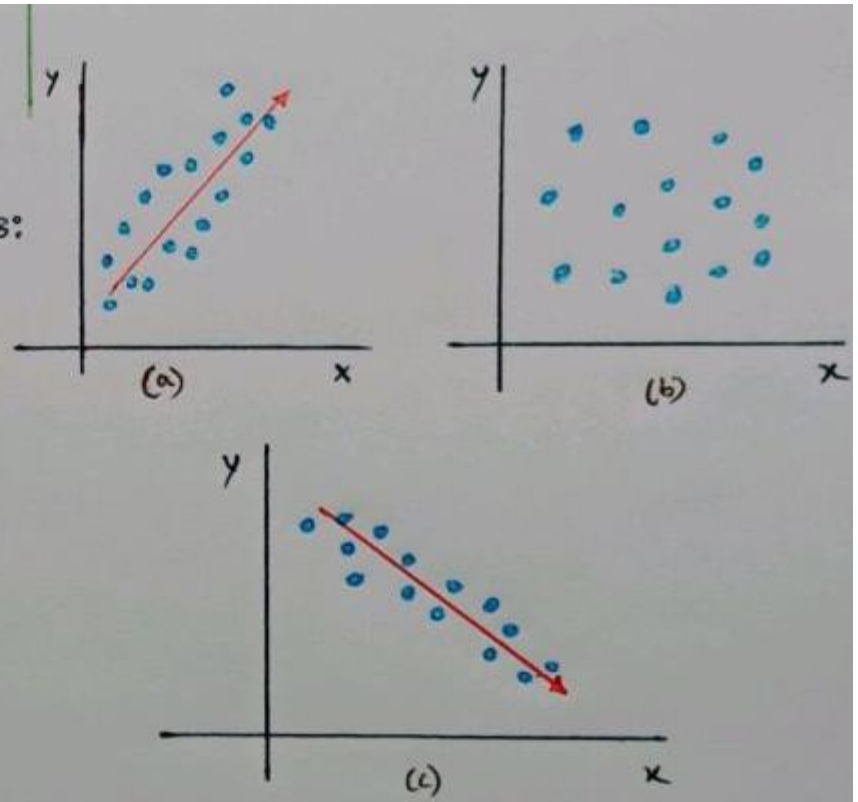
$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Sample

Correlation

Estimates the depth of the relationship between variables.

$$\text{Correlation, } \rho(x,y) = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}$$

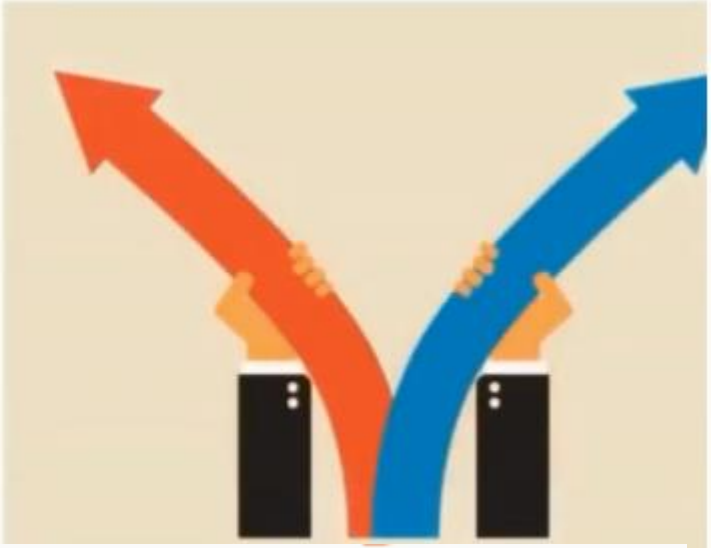


(a) Positive Covariance (b) zero covariance nearly

(c) Negative covariance

Variance , Covariance and correlation

Average of the squared difference of the data from the Mean.



$$\text{Variance, } S_x^2 = \frac{\sum (x - \bar{x})^2}{(N - 1)}$$

$$\text{Covariance, } S_{xy}^2 = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{(N - 1)}$$

Variance of X with respect to Y.

Pearson Correlation Coefficient

$$r = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{(N - 1) * \sigma_x * \sigma_y} = \frac{\text{Covar}(x, y)}{\sigma_x * \sigma_y}$$

Covariance

	Height X	Weight Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})*(Y-\bar{Y})$
	160	130	-15.625	-40.625	634.7656
	170	150	-5.625	-20.625	116.0156
	165	145	-10.625	-25.625	272.2656
	180	190	4.375	19.375	84.76563
	175	175	-0.625	4.375	-2.73438
	190	210	14.375	39.375	566.0156
	185	180	9.375	9.375	87.89063
	180	185	4.375	14.375	62.89063
Mean	**175.625**	**170.625**			**1821.875**
Std Dev	**10.155**	**25.651**			

$$\text{Covariance, } S^2_{xy} = \frac{\sum (x-\bar{x}) * (y-\bar{y})}{(N-1)}$$

$$\text{Covar}(x, y) = \frac{1821.875}{(8-1)}$$

$$\text{Covar}(x, y) = 260.27$$

Positive

Variance Example

Spread in Data

Day	Temperature
1	20
2	21
3	19
4	20
5	21
6	19
7	20
Total	**140**

Mean = 20

Median = 20

Day	Temperature
1	22
2	23
3	21
4	18
5	19
6	17
7	20
Total	**140**

Mean = 20

Median = 20

Day	Temperature
1	12
2	11
3	13
4	20
5	24
6	29
7	31
Total	**140**

Mean = 20

Median = 20

Variance and Standard Deviation

Day	X	$X - \overline{X}$	$(X - \overline{X})^2$
1	20	0	0
2	21	1	1
3	19	-1	1
4	20	0	0
5	21	1	1
6	19	-1	1
7	20	0	0

Average $= 4/7 = 0.57$

Variance, $\sigma^2 = 0.57$

$\sigma = 0.7559$

Mean $= \overline{X} = 20$

Variance and Standard Deviation

Day	X	$X - \overline{X}$	$(X - \overline{X})^2$
1	12	-8	64
2	11	-9	81
3	13	-7	49
4	20	0	0
5	24	4	16
6	29	9	81
7	31	11	121
			412

Average $= 412/7 = 58.857$

Variance, $\sigma^2 = 58.857$

$\sigma = 7.67$

Mean $= \overline{X} = 20$

Variance and Standard Deviation

Day	Temperature
1	20
2	21
3	19
4	20
5	21
6	19
7	20

$\sigma = 0.7559$

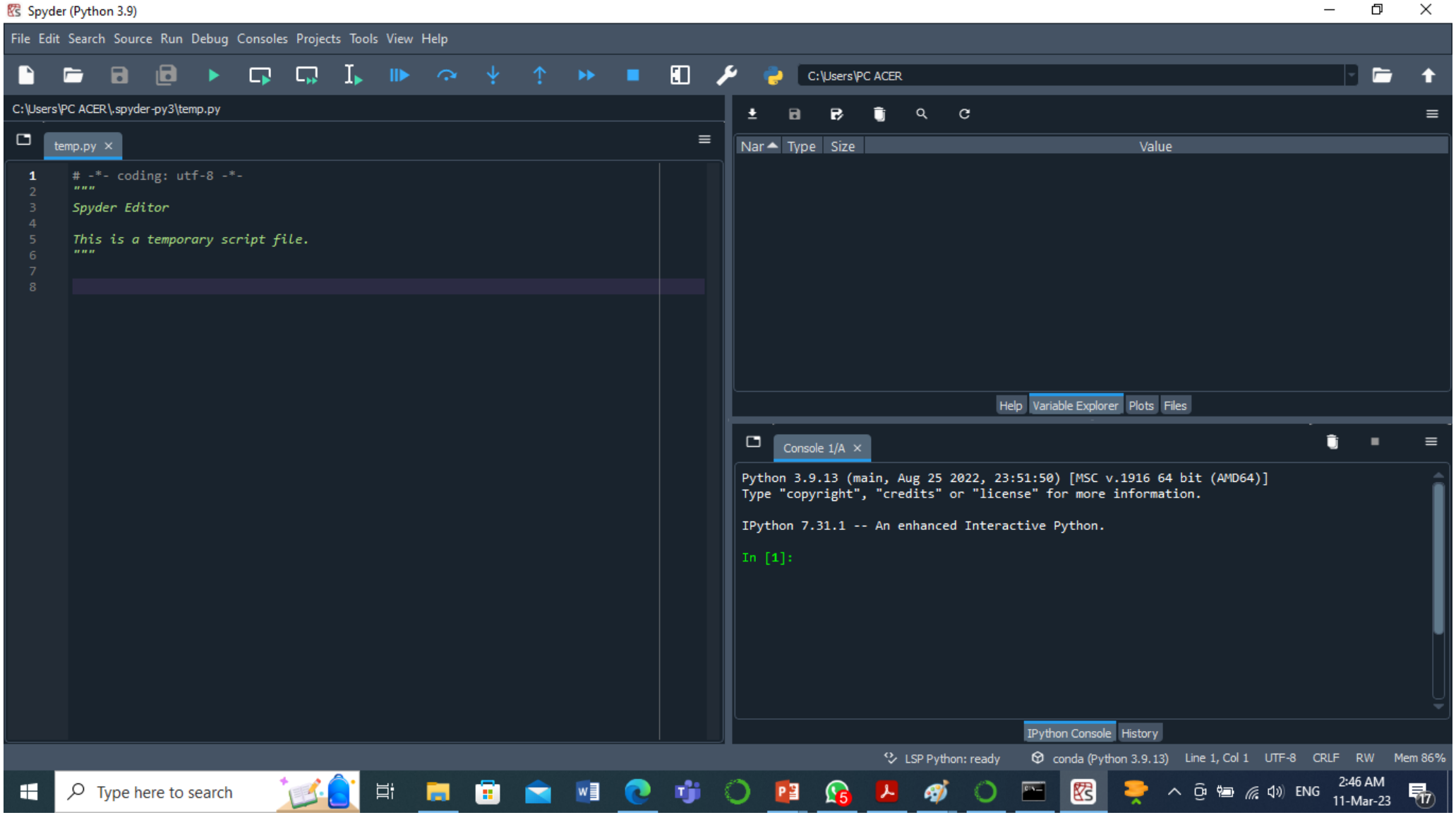
$\text{Mean} = \bar{X} = 20$

Day	Temperature
1	12
2	11
3	13
4	20
5	24
6	29
7	31

$\sigma = 7.67$

$\text{Mean} = \bar{X} = 20$

Read And Slice The Data Using Panda



Spyder (Python 3.9)

File Edit Search Source Run Debug Consoles Projects Tools View Help

D:\1- Sherif folder\2-ACU Courses\Data Science programming\Python-files

D:\1- Sherif folder\2-ACU Courses\Data Science programming\Python-files\Read Data.py

Read Data.py* × loan_small.csv ×

```
1  # -*- coding: utf-8 -*-
2  """
3  Spyder Editor
4
5  This is a temporary script file.
6  """
7
8  import pandas as pd
9
10 dataset=pd.read_csv('loan_small.csv')
11
12
```

Name	Type	Size	Value
dataset	DataFrame	(16, 7)	Column names: Loan_ID, Gender, ApplicantIncome, CoapplicantIncome, Loa ...

Help Variable Explorer Plots Files

Console 1/A ×

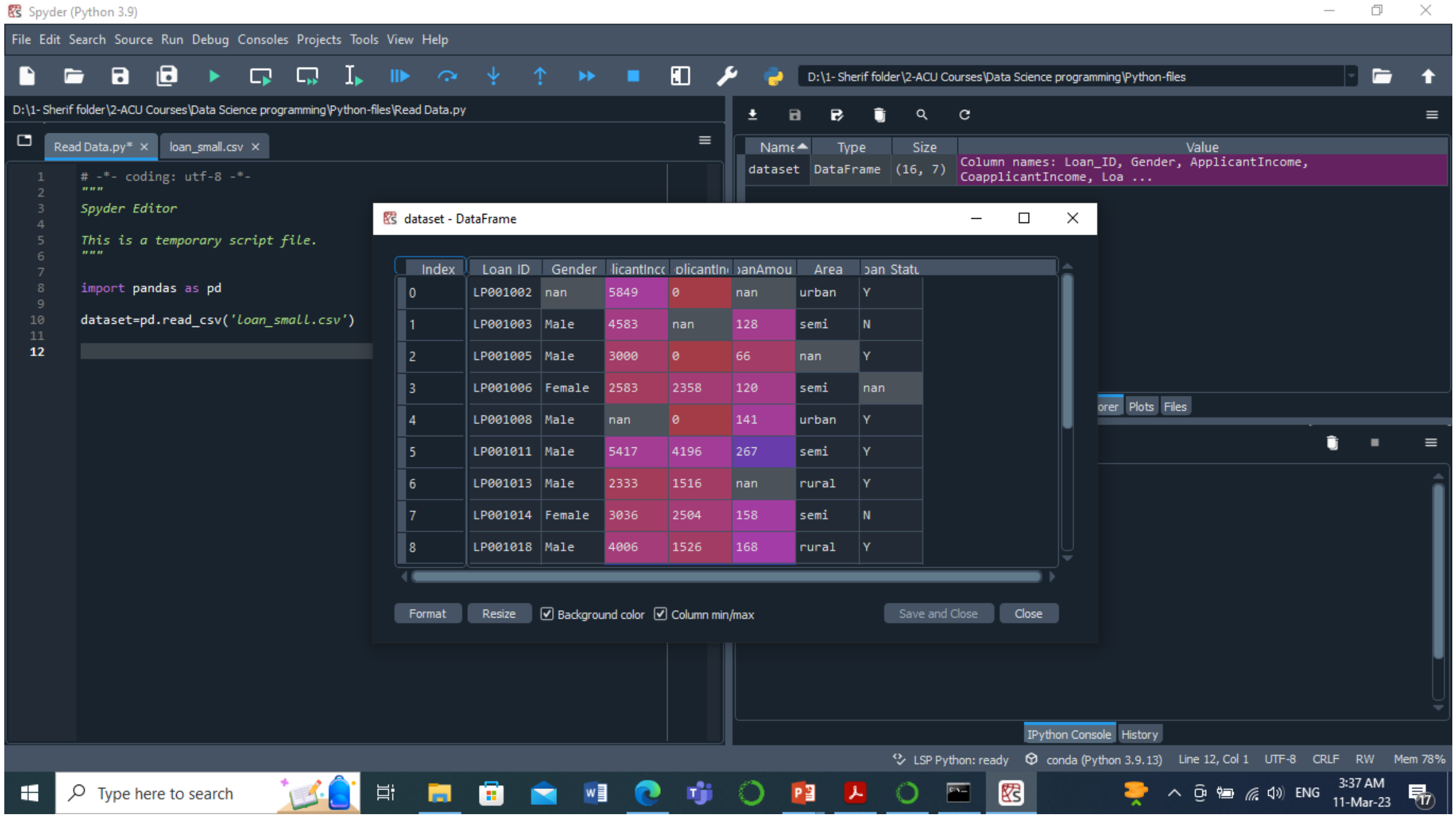
```
In [4]:
```

IPython Console History

LSP Python: ready conda (Python 3.9.13) Line 12, Col 1 UTF-8 CRLF RW Mem 77%

Type here to search

ENG 3:36 AM 11-Mar-23


Spyder (Python 3.9)
File Edit Search Source Run Debug Consoles Projects Tools View Help
D:\1- Sherif folder\2-ACU Courses\Data Science programming\Python-files
D:\1- Sherif folder\2-ACU Courses\Data Science programming\Python-files\Read Data.py
Read Data.py* ×
loan_small.csv ×
1 # -*- coding: utf-8 -*-
2 """
3 Spyder Editor
4
5 This is a temporary script file.
6 """
7
8 import pandas as pd
9
10 dataset=pd.read_csv('loan_small.csv')
11
12
Name Type Size Value
dataset DataFrame (16, 7) Column names: Loan_ID, Gender, ApplicantIncome, CoapplicantIncome, Loa ...
dataset - DataFrame
Index Loan ID Gender licantInc plicantIn anAmou Area an Statu
0 LP001002 nan 5849 0 nan urban Y
1 LP001003 Male 4583 nan 128 semi N
2 LP001005 Male 3000 0 66 nan Y
3 LP001006 Female 2583 2358 120 semi nan
4 LP001008 Male nan 0 141 urban Y
5 LP001011 Male 5417 4196 267 semi Y
6 LP001013 Male 2333 1516 nan rural Y
7 LP001014 Female 3036 2504 158 semi N
8 LP001018 Male 4006 1526 168 rural Y
Format
Resize
Background color
Column min/max
Save and Close
Close
orer Plots Files
IPython Console History
LSP Python: ready
conda (Python 3.9.13)
Line 12, Col 1
UTF-8
CRLF
RW
Mem 78%
Type here to search
ENG
3:37 AM
11-Mar-23