

Introduction to Descriptive Statistics



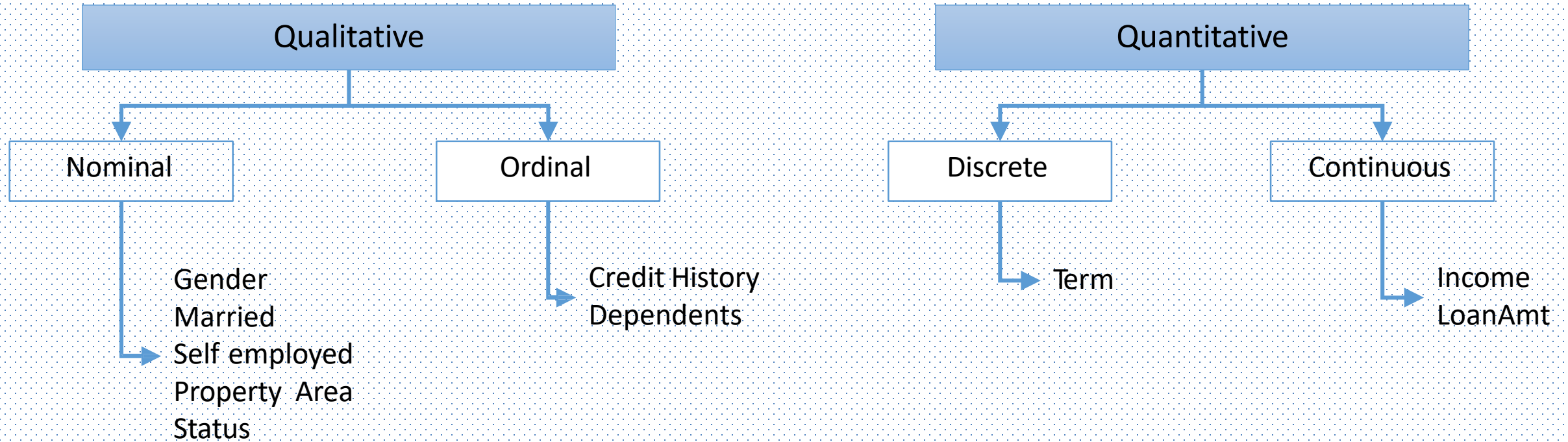
Eslam Ahmed

Software Engineer

Understanding Data, Variables/Features

Understanding The Variables Using a Dataset

Loan_ID	Gender	Married	Dependents	Self_Employed	Income	LoanAmt	Term	CreditHistory	Property_Area	Status
LP001002	Male	No	0	No	\$5,849.00		60	1	Urban	Y
LP001003	Male	Yes	1	No	\$4,583.00	\$128.00	120	1	Rural	N
LP001005	Male	Yes	0	Yes	\$3,000.00	\$66.00	60	1	Urban	Y
LP001006	Male	Yes	2	No	\$2,583.00	\$120.00	60	1	Urban	Y



Understanding The Variables Using a Dataset

Loan_ID	Gender	Married	Dependents	Self_Employed	Income	LoanAmt	Term	CreditHistory	Property_Area	Status
LP001002	Male	No	0	No	\$5,849.00		60	1	Urban	Y
LP001003	Male	Yes	1	No	\$4,583.00	\$128.00	120	1	Rural	N
LP001005	Male	Yes	0	Yes	\$3,000.00	\$66.00	60	1	Urban	Y
LP001006	Male	Yes	2	No	\$2,583.00	\$120.00	60	1	Urban	Y

Data Type

- **Predictor/Independent**

- Gender
- Married
- Dependents
- Self_Employed
- Income
- LoanAmt
- Term
- CreditHistory
- PropertyArea

- **Target/Dependent**

- Status

- **Character/String**

- Gender
- Married
- Self_Employed
- Property_Area
- Status

- **Numeric**

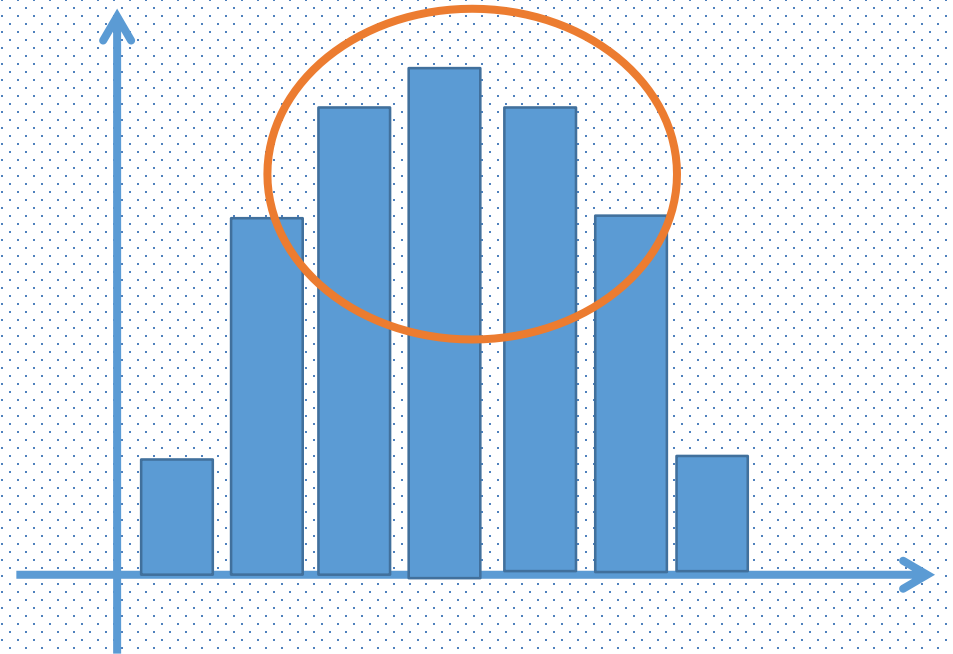
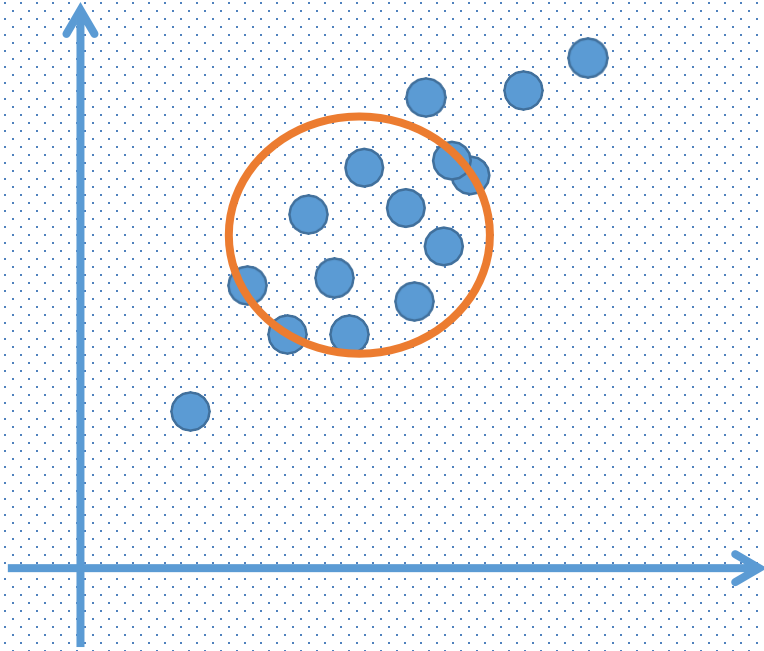
- Dependents
- Income
- LoanAmt
- Term
- CreditHistory

Central Tendency of Data

Central Tendency

Single value that attempts to describe the whole data using a central point or central location of the data.

Central Tendency of Data



Central Tendency

- Mean
- Median
- Mode
- Others – Geometric mean, Harmonic Mean, Weighted Arithmetic Mean

Mean

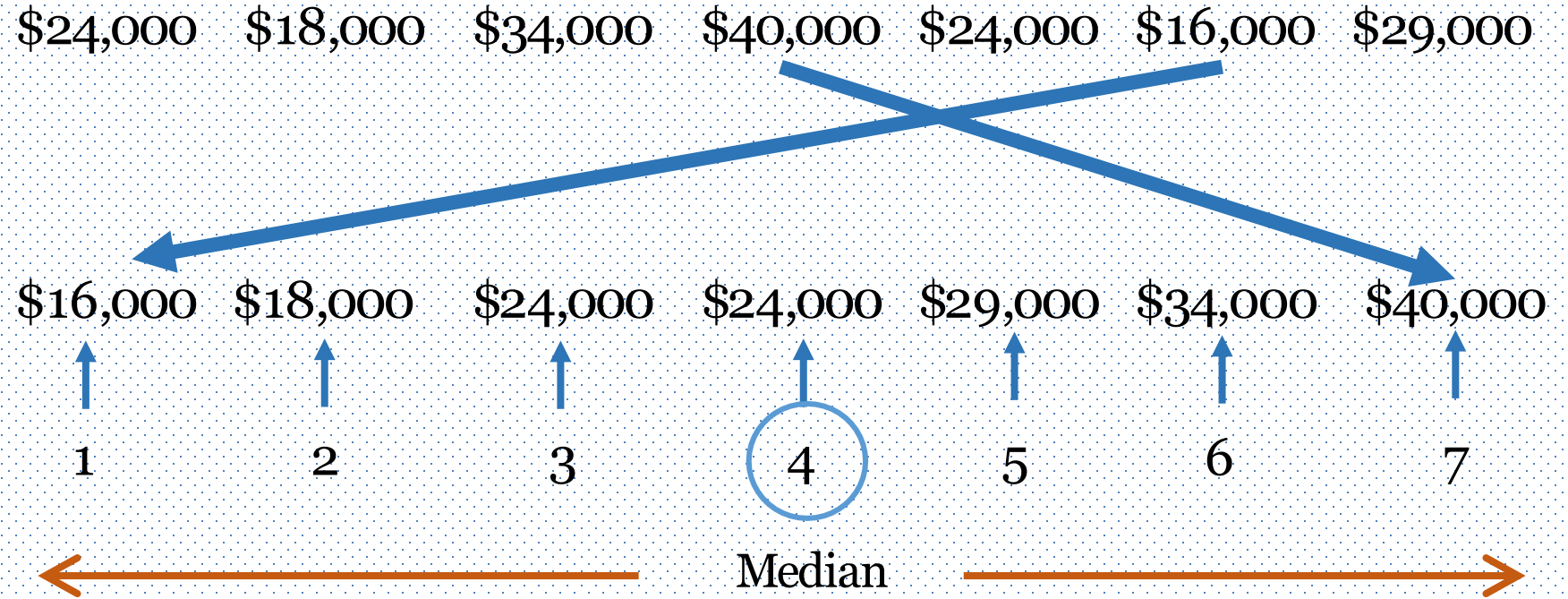
Applicant	Loan Amount
Jitesh	\$ 24,000
John	\$ 18,000
Frans	\$ 34, 000
Danny	\$ 40,000
Cecile	\$ 24,000
Scott	\$ 16,000
Alex	\$ 29,000

$$\begin{aligned}\text{Mean} &= \frac{24000 + 18000 + 34000 + 40000 + 24000 + 16000 + 29000}{7} \\ &= \frac{151000}{7}\end{aligned}$$

$$\text{Mean} = \$25,167$$

Median

Applicant	Loan Amount
Jitesh	\$ 24,000
John	\$ 18,000
Frans	\$ 34, 000
Danny	\$ 40,000
Cecile	\$ 24,000
Scott	\$ 16,000
Alex	\$ 29,000



Median = \$24,000

Mode

Applicant	Loan Amount
Jitesh	\$ 24,000
John	\$ 18,000
Frans	\$ 34, 000
Danny	\$ 40,000
Cecile	\$ 24,000
Scott	\$ 16,000
Alex	\$ 29,000

Mode = \$24,000

Outliers



Experience	Salary
1	\$ 3,725
2	\$ 4,155
3	\$ 4,627
4	\$ 5,147
5	\$ 5,718
6	\$ 6,347
7	\$ 7,039
8	\$ 7,210
9	\$ 7,423
10	\$ 19,000
11	\$ 8,369
12	\$ 8,810
13	\$ 8,940
14	\$ 9,200
15	\$ 9,458

Effect of Outliers

Experience	Salary
1	\$ 3,725
2	\$ 4,155
3	\$ 4,627
4	\$ 5,147
5	\$ 5,718
6	\$ 6,347
7	\$ 7,039
8	\$ 7,210
9	\$ 7,423
10	\$ 7,556
11	\$ 8,369
12	\$ 8,810
13	\$ 8,940
14	\$ 9,200
15	\$ 9,458

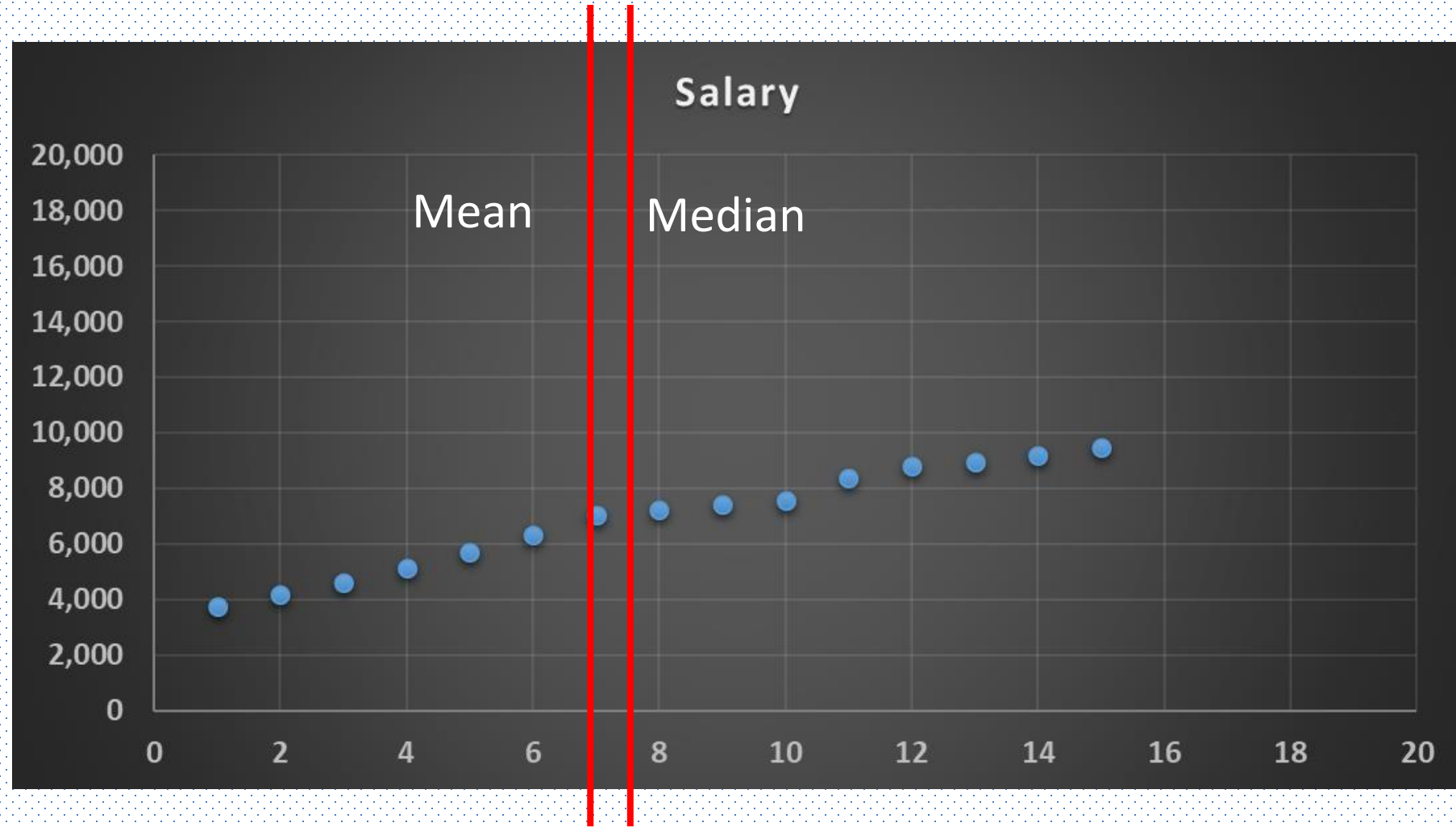
\$6,915 ← Mean → \$7,678

\$7,200 ← Median → \$7,200

Experience	Salary
1	\$ 3,725
2	\$ 4,155
3	\$ 4,627
4	\$ 5,147
5	\$ 5,718
6	\$ 6,347
7	\$ 7,039
8	\$ 7,210
9	\$ 7,423
10	\$ 19,000
11	\$ 8,369
12	\$ 8,810
13	\$ 8,940
14	\$ 9,200
15	\$ 9,458

Measure of Dispersion

Central Tendency



Spread in Data



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

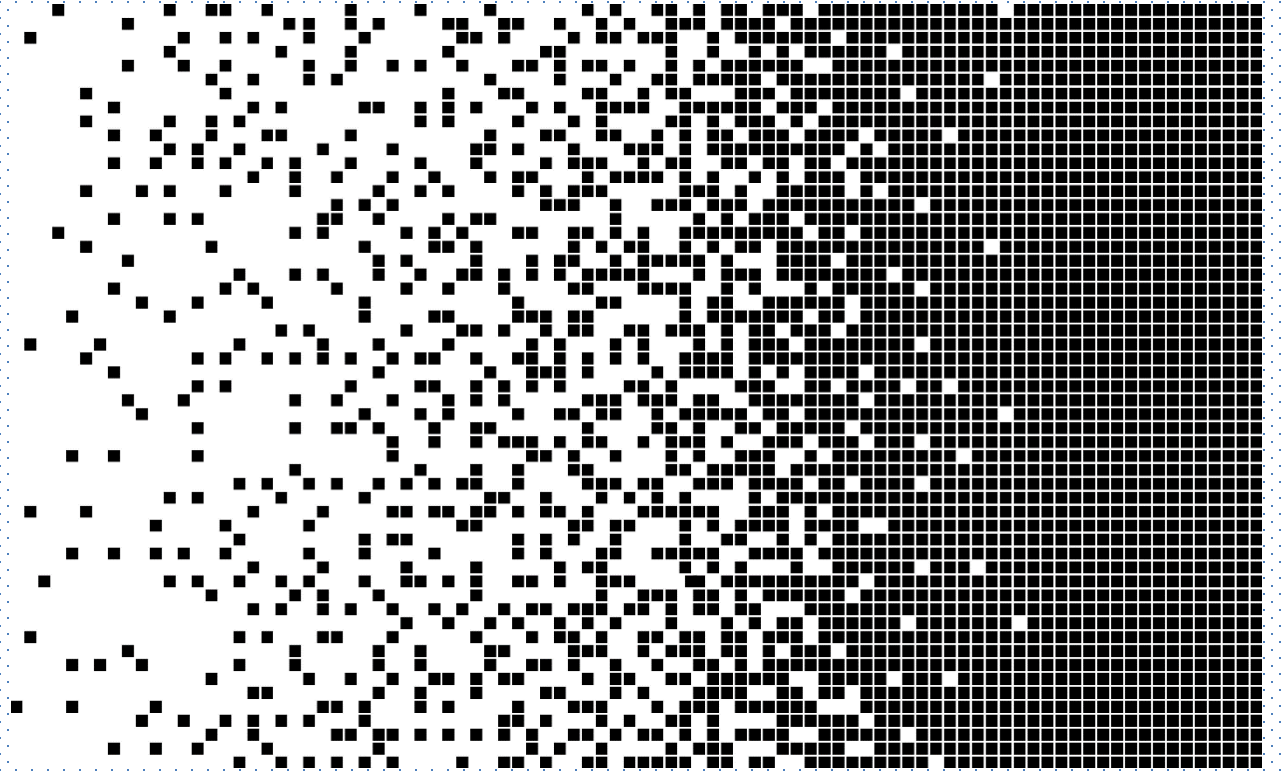
Day	Temperature
1	12
2	11
3	13
4	20
5	24
6	29
7	31
Total	140

Mean = 20

Median = 20

Measure of Dispersion

- Variance
- Standard Deviation
- Percentile
- Range
- Interquartile range



Variance and Standard Deviation

Day	X	$X - \bar{X}$	$(X - \bar{X})^2$
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

$$\text{Average} = 4/7 = 0.57$$

$$\text{Variance, } \sigma^2 = 0.57$$

$$\sigma = 0.7559$$

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

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

$$\sigma = 7.67$$

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

What is Percentile?

The value below which a given percentage of observations in a group of observations falls...

– Wikipedia

Percentile

- Arrange the data in an order
- Calculate the percentage of observations or data points below a particular value.

What is the 80th Percentile Observation?

Total Observations * 0.8

$$15 * 0.8 = 12$$



Row Number	Salary
1	\$ 3,725
2	\$ 4,155
3	\$ 4,627
4	\$ 5,147
5	\$ 5,718
6	\$ 6,347
7	\$ 7,039
8	\$ 7,210
9	\$ 7,423
10	\$ 7,556
11	\$ 8,369
12	\$ 8,810
13	\$ 8,940
14	\$ 9,200
15	\$ 9,458

Range

Difference between the highest and lowest value...

Range

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

Mean = 20
Range = 2

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

Mean = 20
Range = 6

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

Mean = 20
Range = 20

Inter Quartile Range (IQR)

1st Quartile



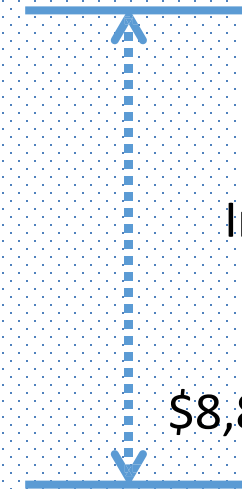
Median



3rd Quartile



Row Number	Salary
1	\$ 3,725
2	\$ 4,155
3	\$ 4,627
4	\$ 5,147
5	\$ 5,718
6	\$ 6,347
7	\$ 7,039
8	\$ 7,210
9	\$ 7,423
10	\$ 7,556
11	\$ 8,369
12	\$ 8,810
13	\$ 8,940
14	\$ 9,200
15	\$ 9,458



$Q3 - Q1$
Inter Quartile Range
IQR

$$\$8,810 - \$5,147 = \$3,663$$

How to Show Numerical Data?

Visualize Numerical Data

- Frequency Table
- Histogram
- Bar Chart
- Boxplot

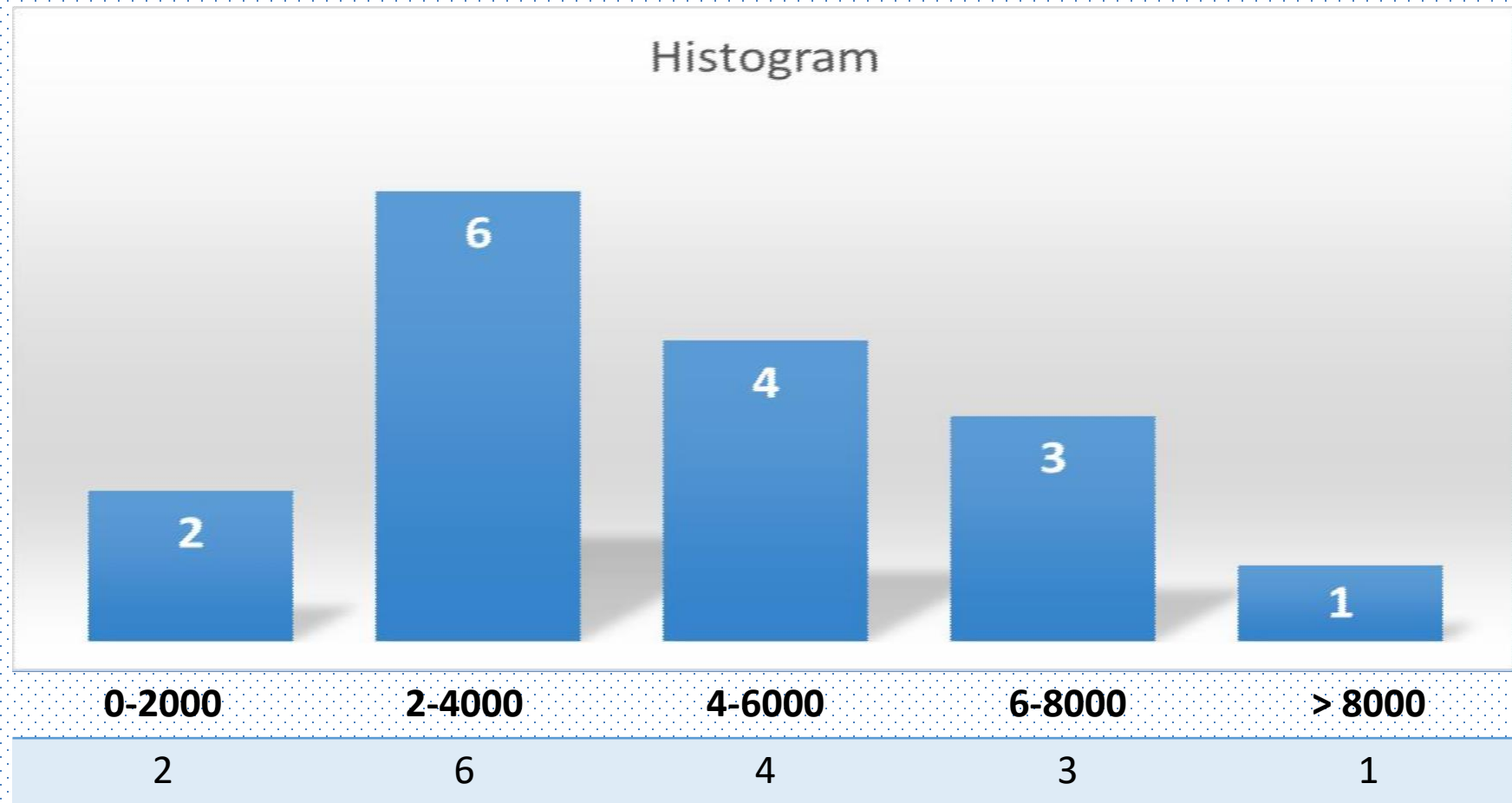
Frequency Table

1223
3434
4545
6798
2311
4321
5600
10345
900
2687
3450
6700
2340
3600
5632
7900

0-2000	2-4000	4-6000	6-8000	> 8000
1223	3434	4545	6798	10345
900	2311	4321	6700	
	2687	5600	7900	
	3450	5632		
	2340			
	3600			
2	6	4	3	1

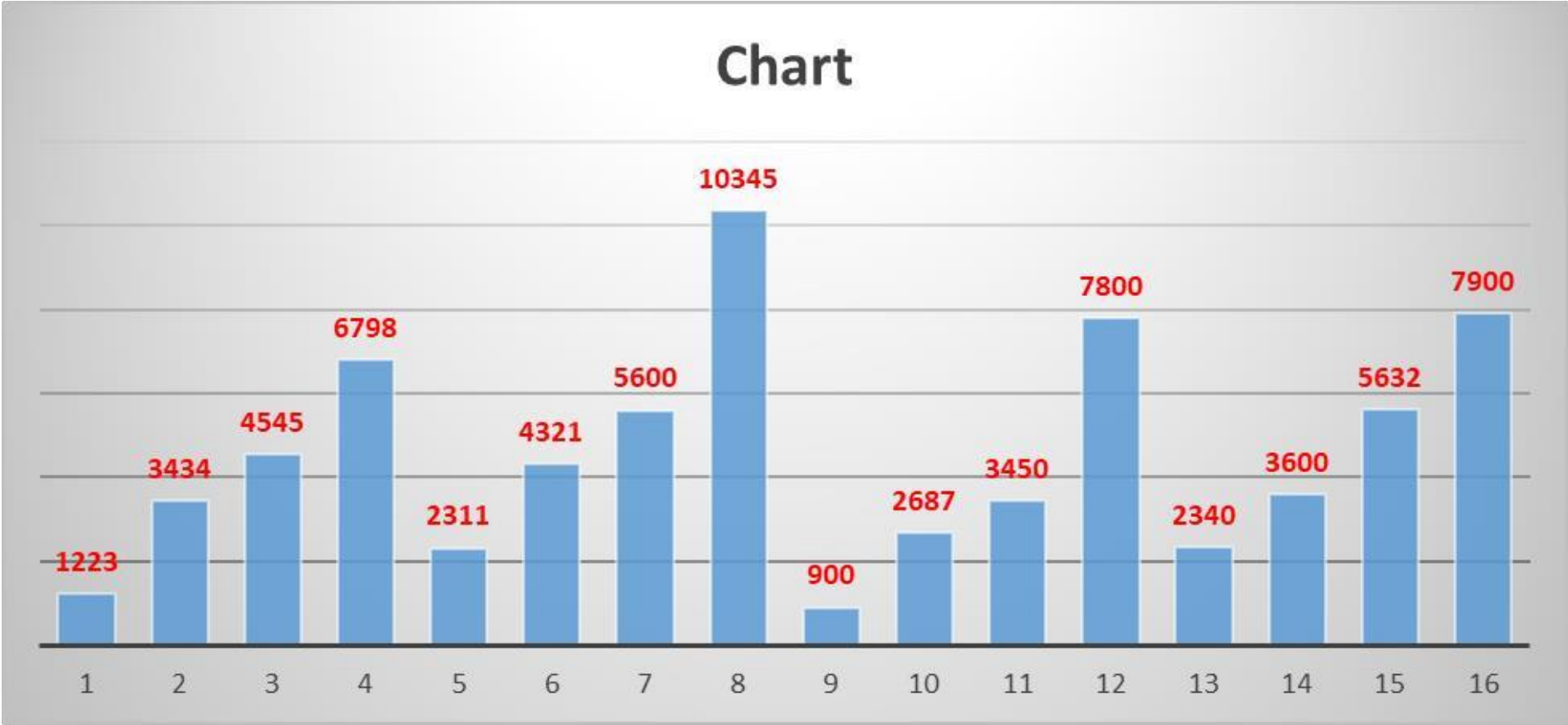
Histogram

1223
3434
4545
6798
2311
4321
5600
10345
900
2687
3450
6700
2340
3600
5632
7900

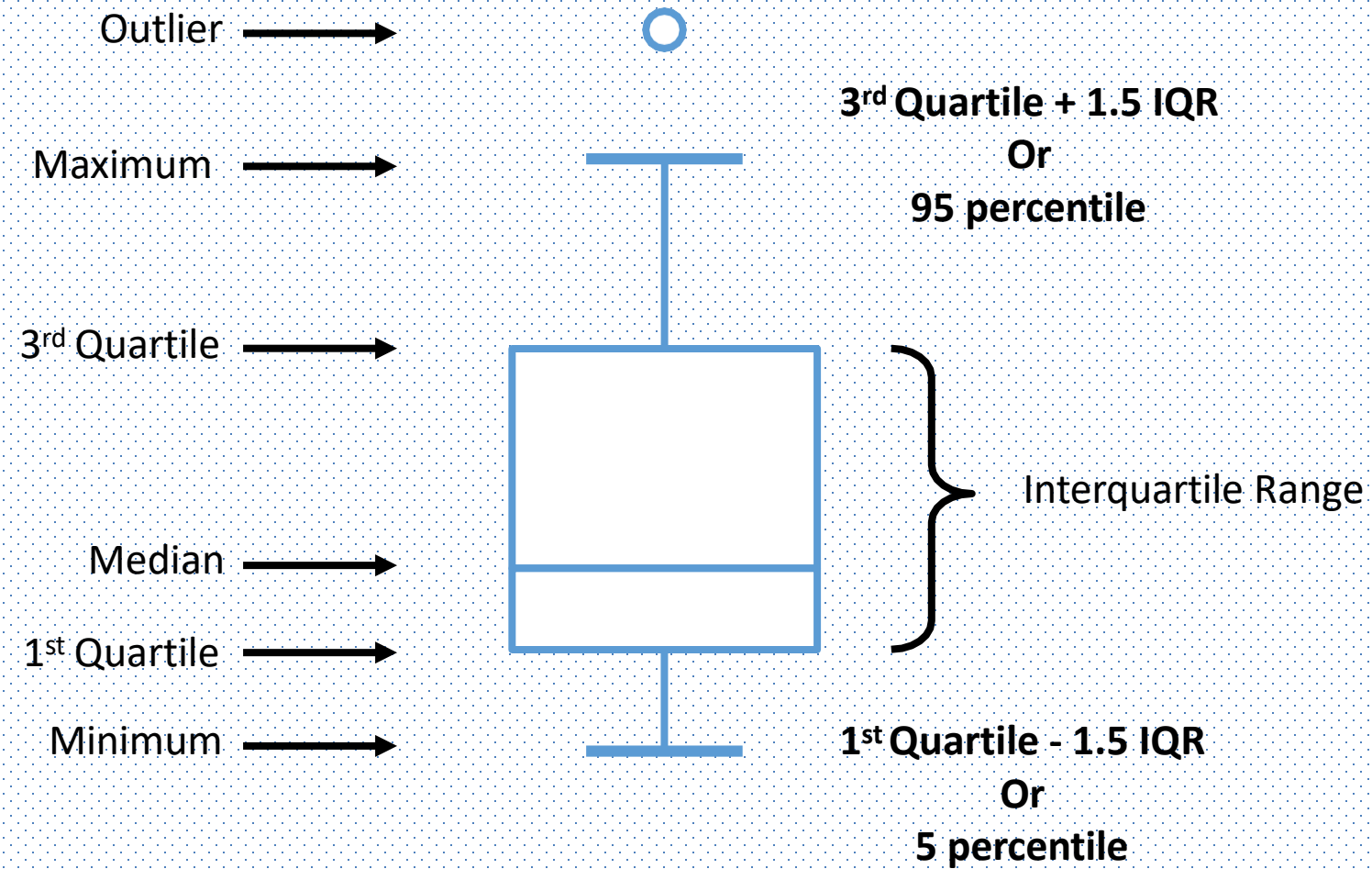


Bar Chart

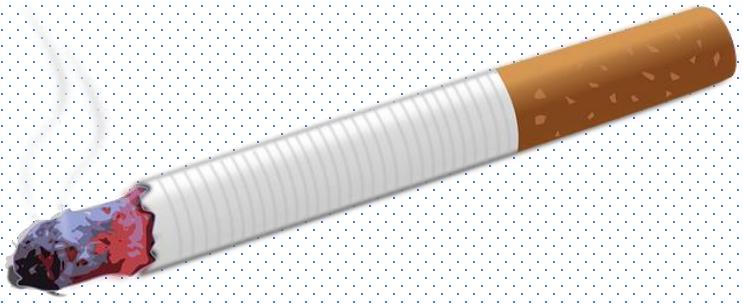
1223
3434
4545
6798
2311
4321
5600
10345
900
2687
3450
6700
2340
3600
5632
7900



Box Plot



Correlation



Number of cigarettes smoked

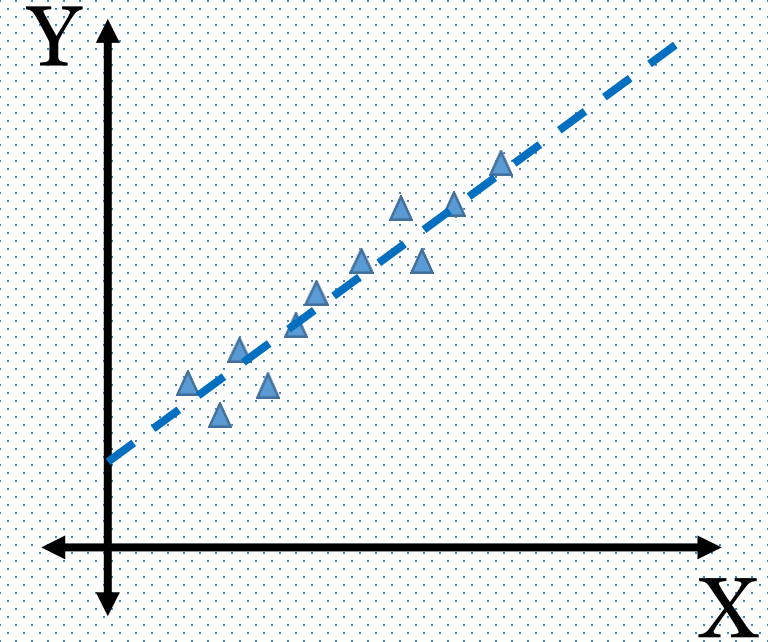
Stress Level

Statistically Correlated

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

Pearson Correlation
Coefficient

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



Correlation Coefficient

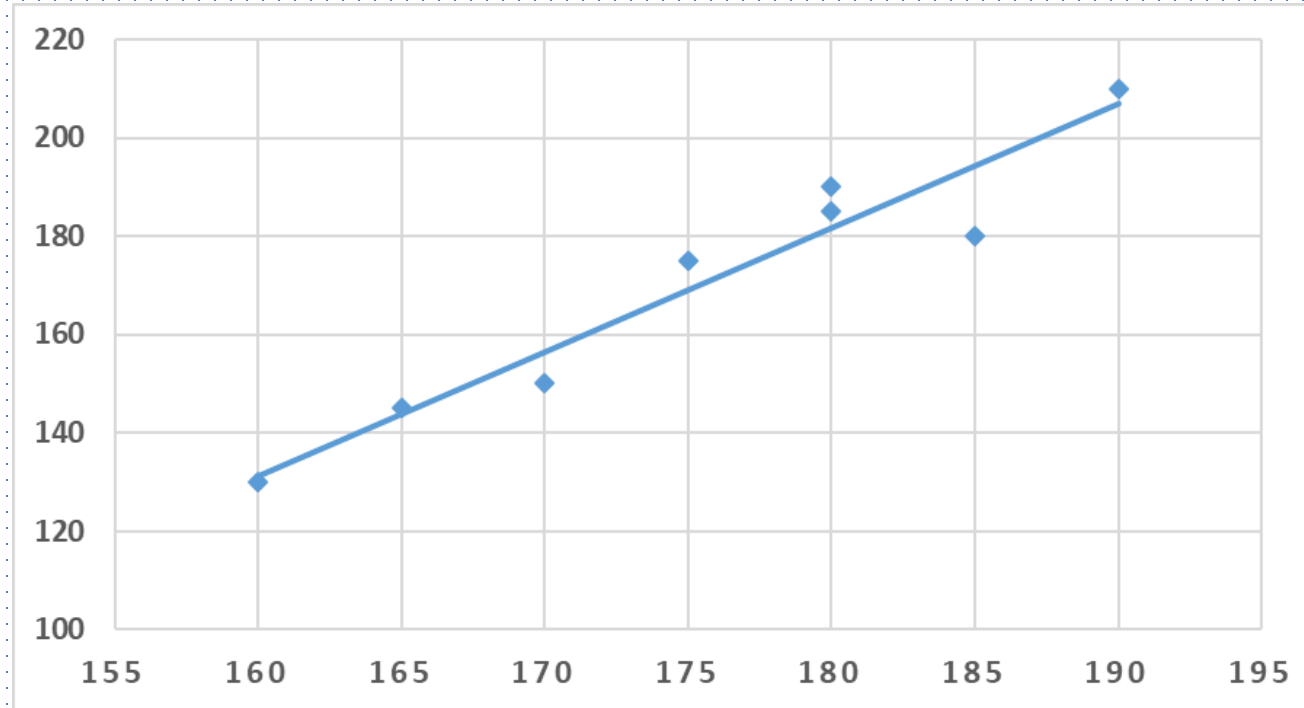
	Height X	Weight Y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) * (y - \bar{y})$
	160	130	-15.625	-40.625	634.7656
	170	150	-5.625	-20.625	116.0156
	165	145	-10.625	-25.625	272.2656
	180	190	4.375	19.375	84.76563
	175	175	-0.625	4.375	-2.73438
	190	210	14.375	39.375	566.0156
	185	180	9.375	9.375	87.89063
	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

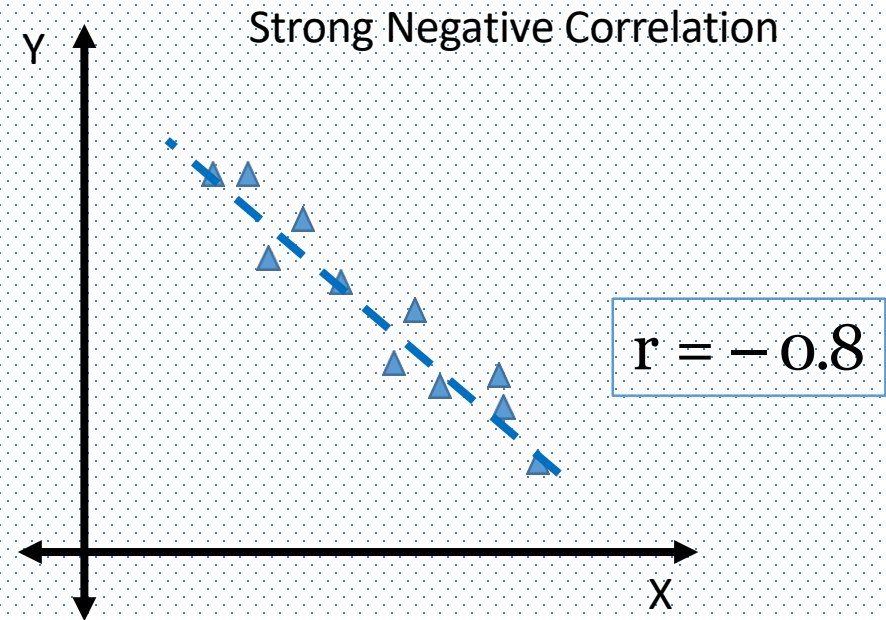
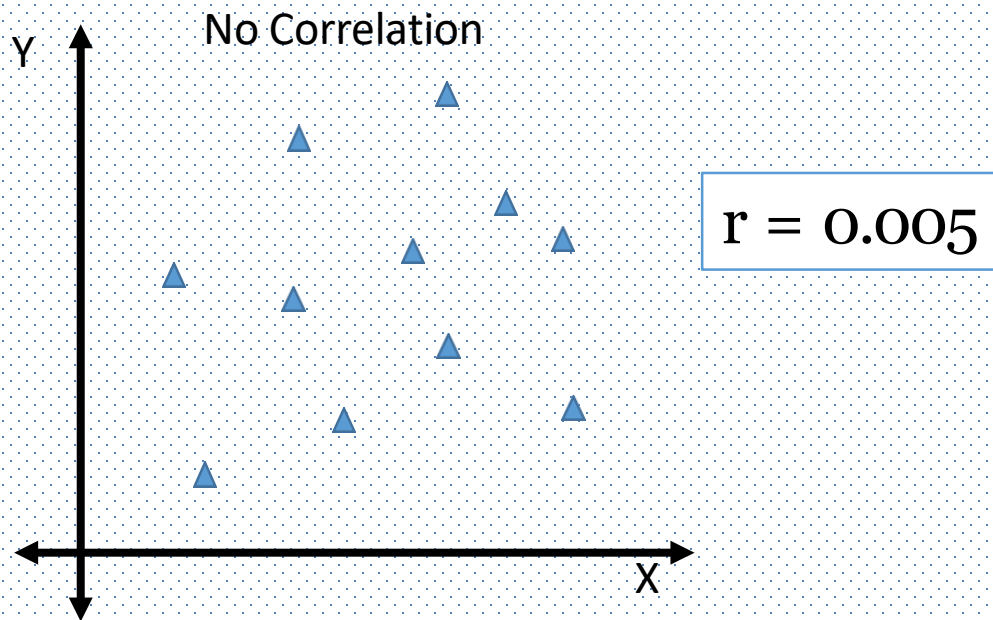
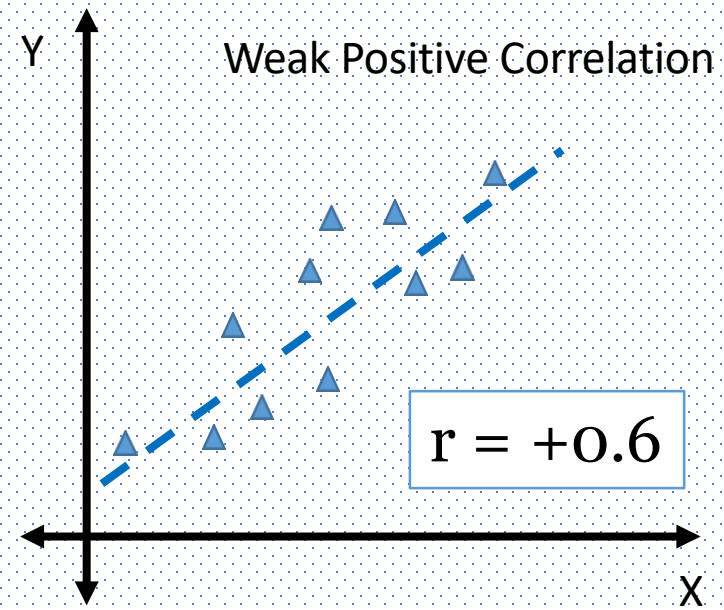
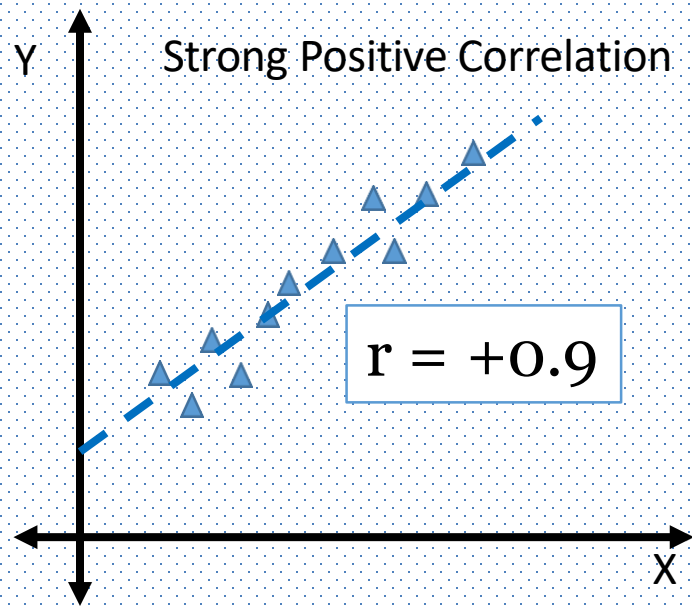


Scatter Plot

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

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

$$r = 0.96$$



Covariance

Variance

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

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

Variance of X with
respect to X.

Covariance

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

Variance of X with
respect to Y.

Covariance

Pearson Correlation
Coefficient

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

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

Variance of X with
respect to 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			

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

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

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

Covariance

- Non-Standardised method of correlation
- Can be positive or negative

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

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

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

Covariance Matrix

	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			

	X	Y
X	Covariance (x, x)	Covariance(x, y)
Y	Covariance (y, x)	Covariance (y, y)

Covariance Matrix

	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			

	X	Y
X	Variance(x)	Covariance(x, y)
Y	Covariance (y, x)	Variance (y)

Variance – Covariance Matrix

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

Covariance Matrix

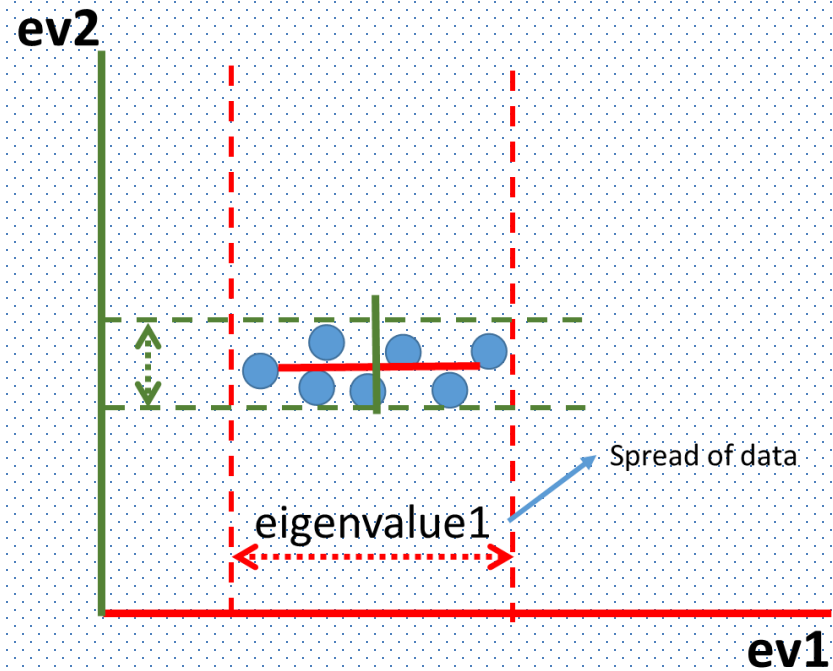
	Height X	Weight Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X}) * (Y - \bar{Y})$
	160	130	-15.625	-40.625	634.7656
	170	150	-5.625	-20.625	116.0156
	165	145	-10.625	-25.625	272.2656
	180	190	4.375	19.375	84.76563
	175	175	-0.625	4.375	-2.73438
	190	210	14.375	39.375	566.0156
	185	180	9.375	9.375	87.89063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

	X	Y
X	103.125	260.27
Y	260.27	710.26

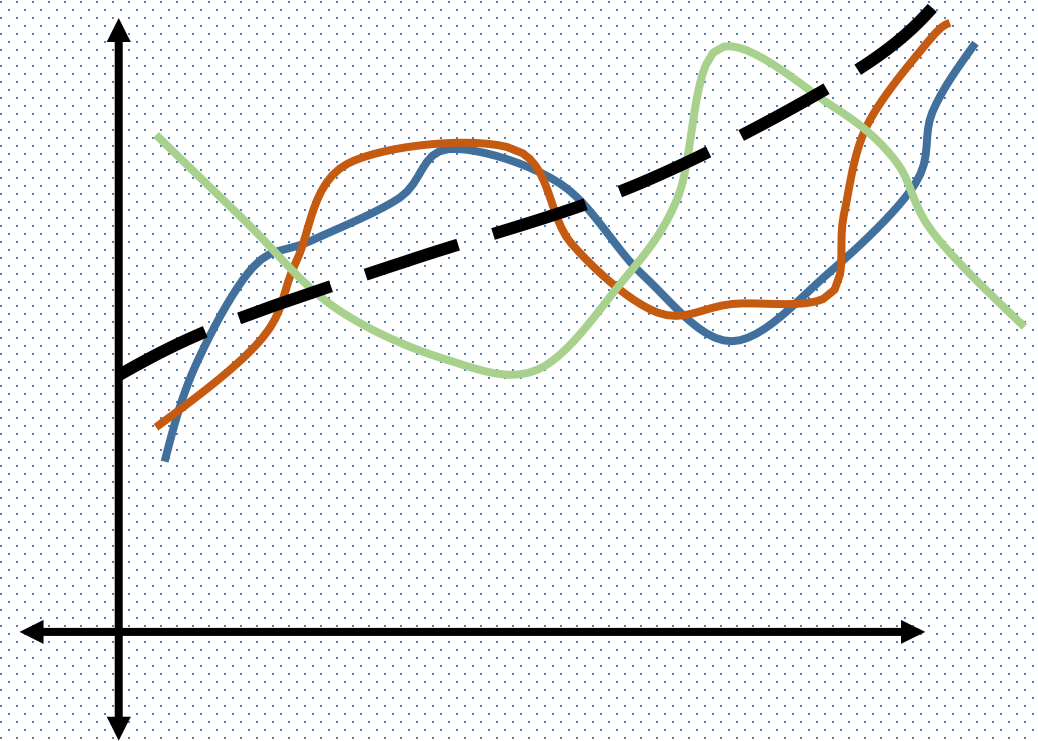
Variance – Covariance Matrix

Covariance Applications

- Using Covariance matrix as Transformation Matrix to get Eigenvectors and EigenValues



- Financial Portfolio Management



Thank You!