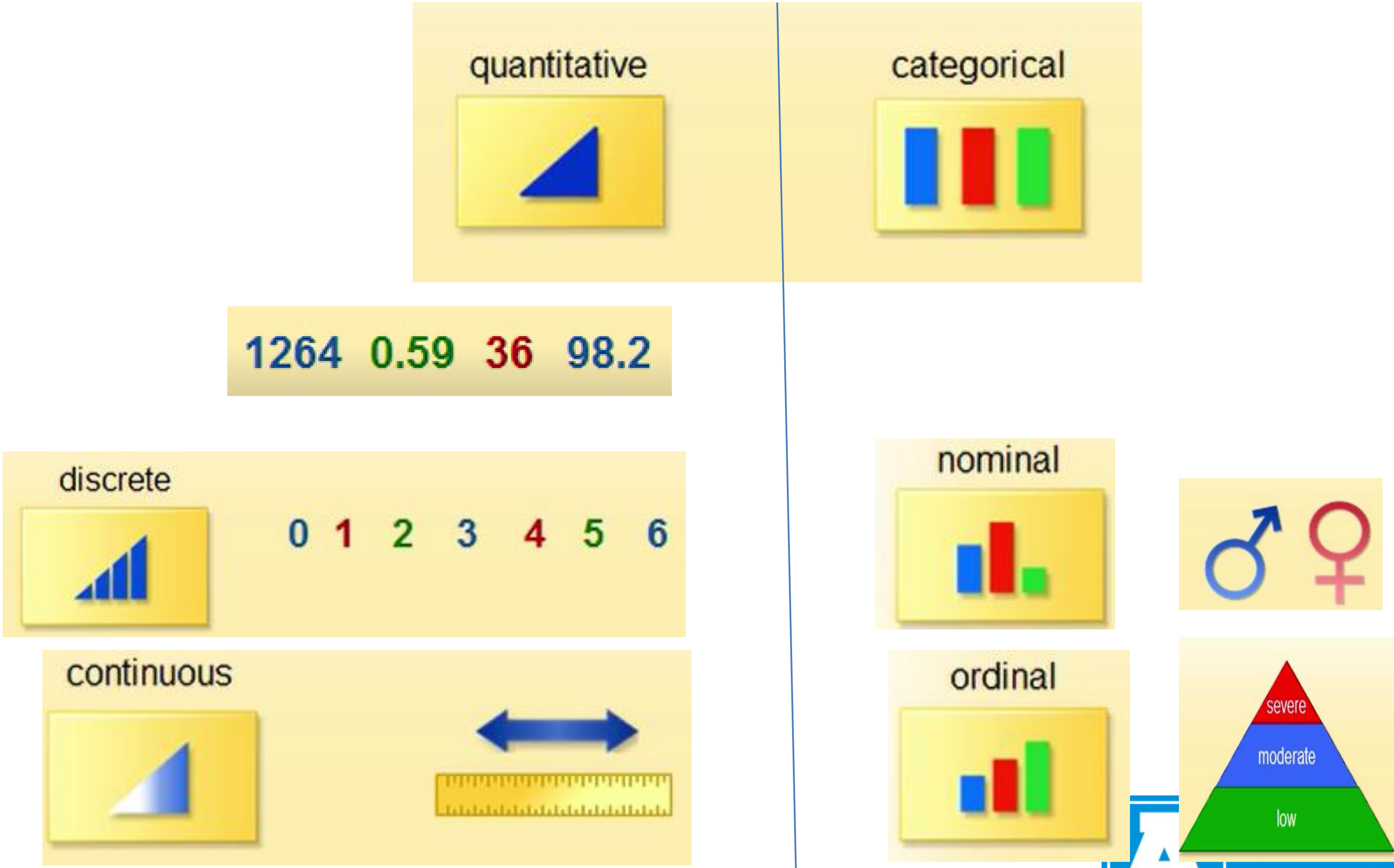
Statistic Fundaments

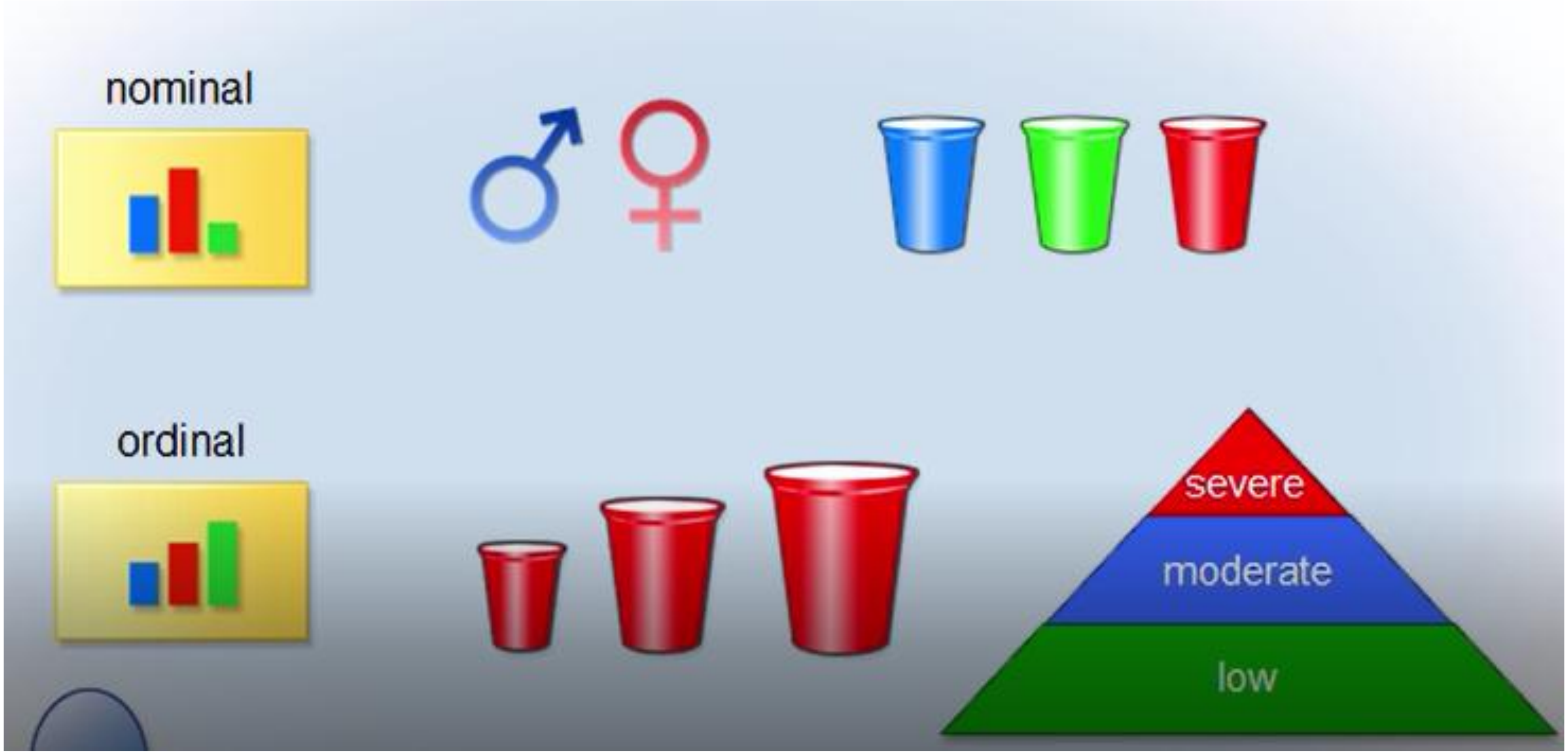
Antrix Academy of Data Science

Introduction to Data

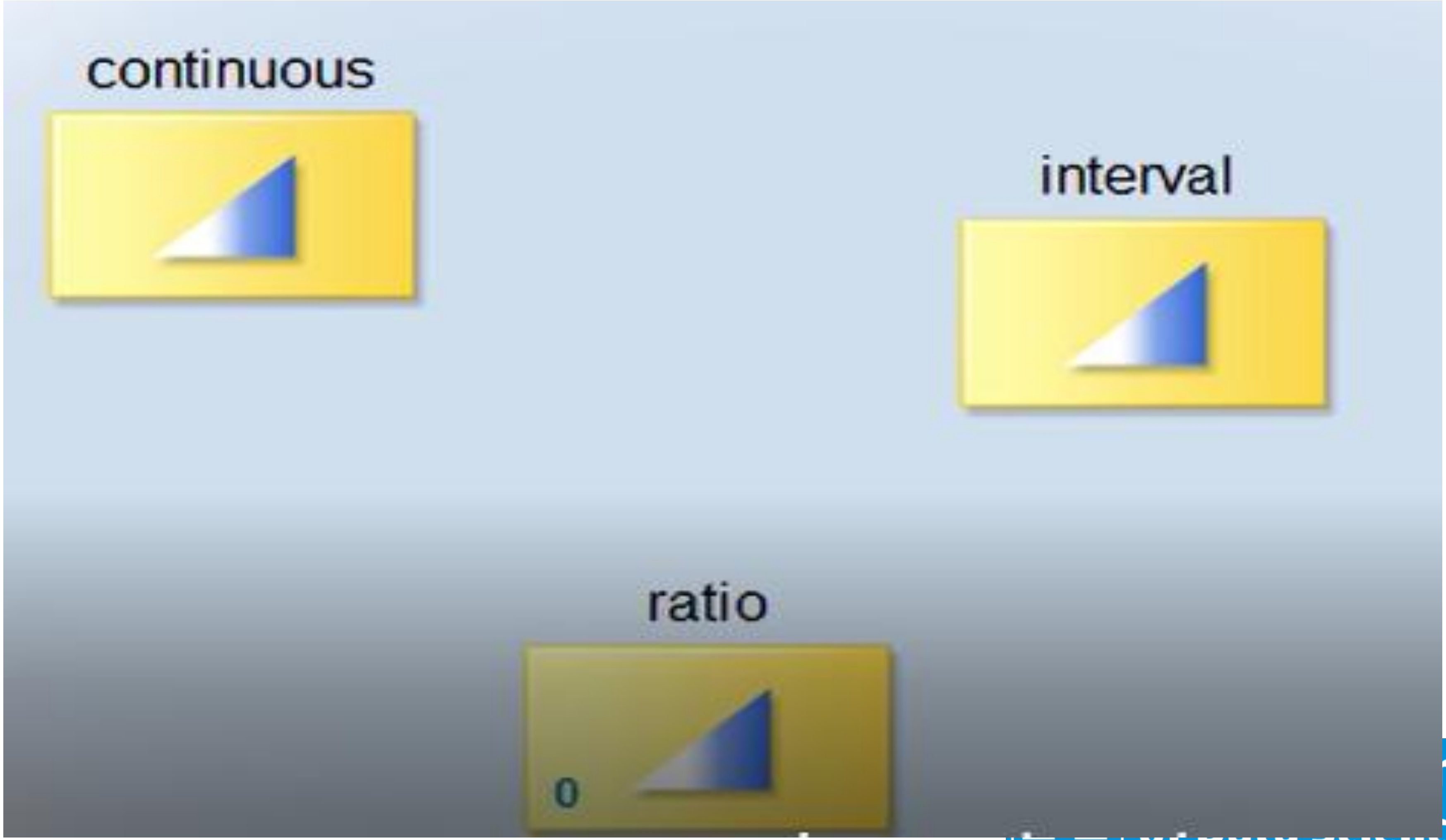


Academy of Data Science

Scale of Measurement



Scale of Measurement...



Univariante Analysis

- Range/Variability
  - Max Value (minus) Min Value
- Central Tendency/Measurement of Data
  - Mean
  - Median
  - Mode
  - Percentile
- Frequency
  - Summary of count of unique values for given variable
  - Can help in getting outliers
- Dispereresion/ Spread
  - Standard Deviation

Range

Range = Maximum value from a list – Minimum value from a list

Range = 16-12 => 4

Age
12
14
15
14
15
12
16

Mean

Mean : Average of values of a list

Mean =97/7 = 13.85

$$\overline{X} = M = \frac{\Sigma X}{n}$$

$$\mu = \frac{\Sigma X}{N}$$

Age
11
14
15
14
15
12
16

Median

Median: Mid value of sorted data of a list

Age
11
14
15
14
15
12
16

Age
11
12
14
14
15
15
16

Antrix Academy of Data Science

Mode

- Mode : a value with highest frequency in a list

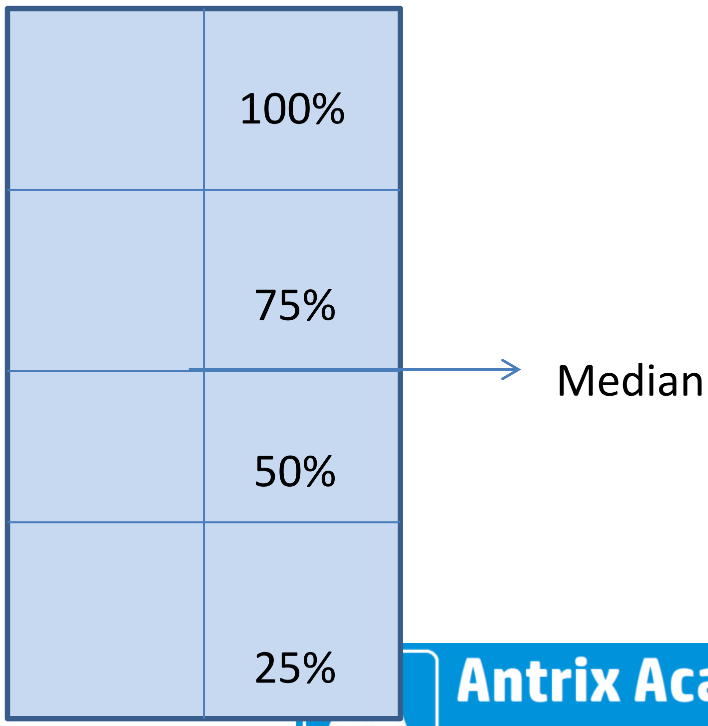
Age
11
14
15
14
15
12
15

Age	Freq. Counts
11	1
12	1
14	2
15	3

Antrix Academy of Data Science

Percentile

- Percentile : generally used to have 4 Quantiles to represent the values of list.



Standard Deviation

- Standard Deviation : Square root of average variance of list data from mean

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

- E.g. Mean is 150 and Standard Deviation is 30

% of Data Range	Value Range	Desc
68%	120-180	Mean +/- (sd)
95%	90-210	Mean +/- 2x(sd)
99.7%	60-240	Mean +/- 3x(sd)

Coefficient of Variance

- It is measure of standard deviation expressed as a percentage of the mean.
- It is way to standardize units of measurement

$$\frac{s}{\overline{X}} \times 100$$

Antrix Academy of Data Science

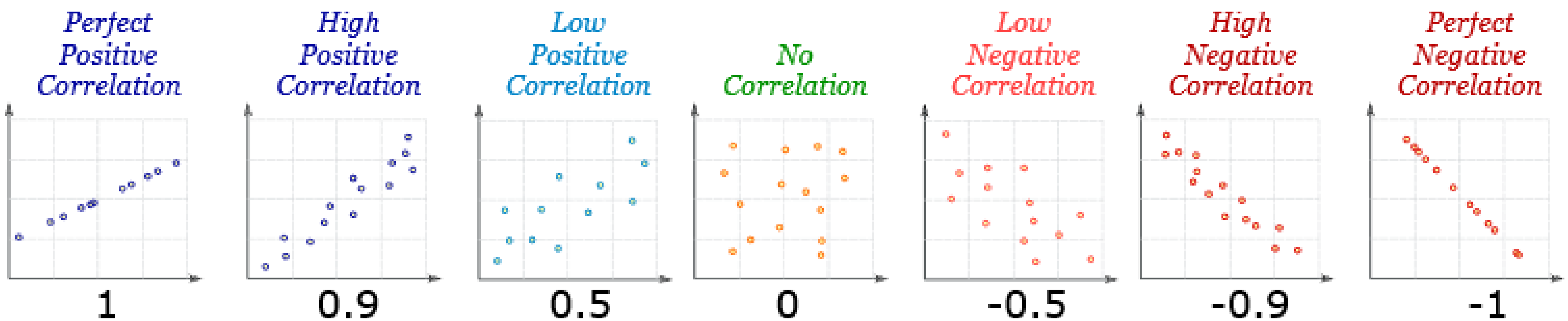
Bivariate Analysis

- Correlation (r)
- Chi-square test
- Linear Regression

Antrix Academy
of Data Science

Correlation

- When two sets of data are strongly linked together we say they have a High Correlation.



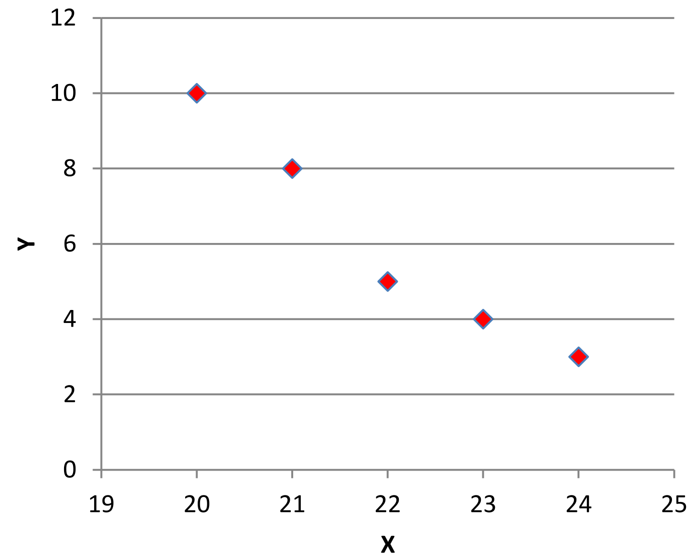
$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

Antrix Academy of Data Science

X	Y	XY	X SQR	Y SQR
100	10	1000	10000	100
110	8	880	12100	64
120	5	600	14400	25
125	4	500	15625	16
130	3	390	16900	9
585	30	3370	69025	214



$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

5	N	=	number of pairs of scores
3370	Σxy	=	sum of the products of paired scores
585	Σx	=	sum of x scores
30	Σy	=	sum of y scores
69025	Σx^2	=	sum of squared x scores
214	Σy^2	=	sum of squared y scores

Correlation : -0.97619

Antrix Academy of Data Science

Chi-Square Test

- Goodness to Fit Test is used to perform hypothesis tests to compare two or more populations
- Null hypothesis => The stated distribution is accurate

Chi-Square

Season	Last Year	% Last Year	Expected Birth	% Historical
Winter	45	22.5%	30	15%
Spring	48	24.0%	40	22%
Summer	55	27.5%	60	30%
Fall	52	26.0%	60	30%
Total	**200**			

Critical Value Comparison :

Here degree of freedom is (4-1)=3
Use 0.05 Column for 5% significant level

P Value = 7.8147

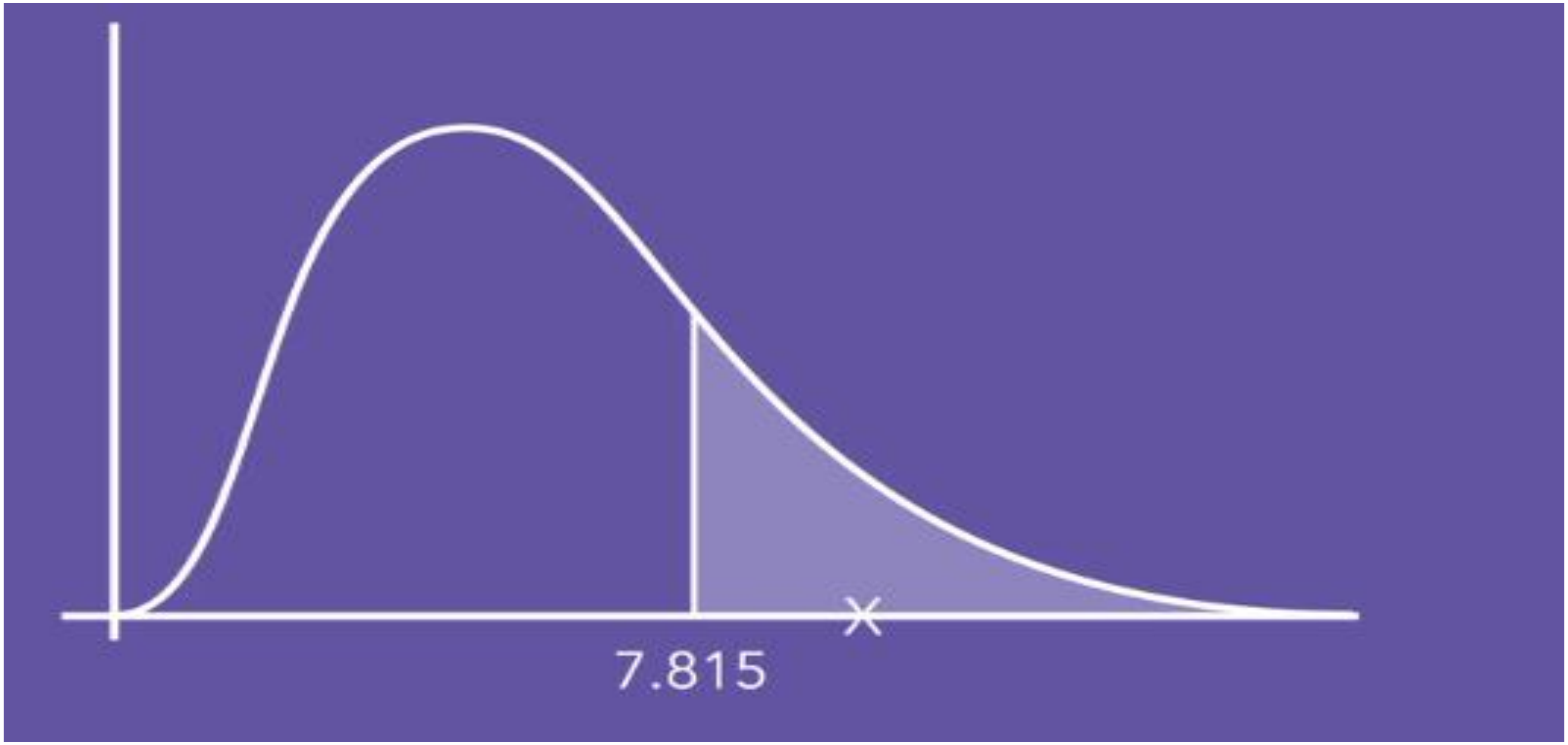
$$X^2_{winter} = \sum \frac{(45-30)^2}{30} = 7.50 \qquad X^2_{summer} = \sum \frac{(55-60)^2}{60} = 0.42$$

$$X^2_{spring} = \sum \frac{(48-50)^2}{50} = 0.08 \qquad X^2_{fall} = \sum \frac{(52-60)^2}{60} = 1.07$$

$$7.50 + 0.08 + 0.42 + 1.07 = 9.07$$

$$X^2 = \sum \frac{(observed - expected)^2}{expected}$$

Antrix Academy of Data Science



- As our values is 9.03 as it is higher than the P value received from chi-square table, therefore we are rejecting the null hypothesis

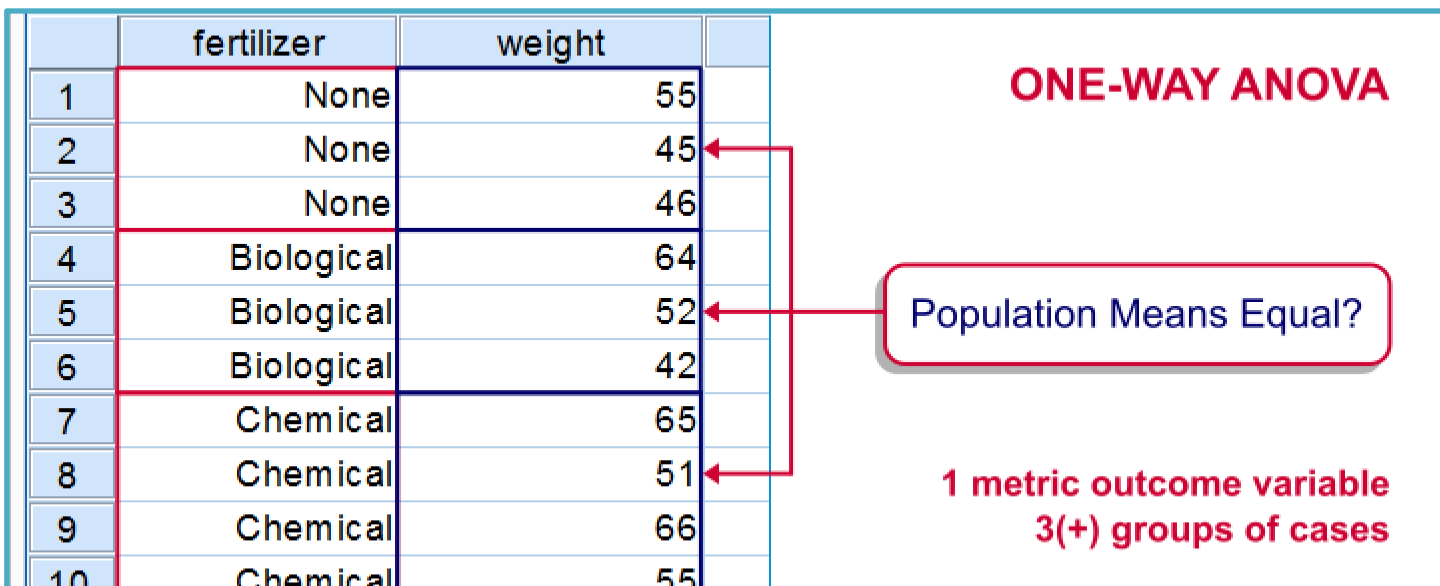
Antrix Academy of Data Science

Multivariate Analytics

- nway ANOVA
- Multiple Linear Regression

ANOVA

- ANOVA -short for “analysis of variance”- is a statistical technique for testing if 3(+) population means are all equal.



Antrix Academy of Data Science

ANOVA...

- one-way ANOVA for comparing **3(+) groups on 1 variable**: do all children from school A, B and C have equal mean IQ scores? *
- Repeated Measures ANOVA for comparing **3(+) variables in 1 group**: is the mean rating for beer A, B and C equal for all people?

SST

Air Mobile	Binge Tech	ComMobile	Data Roam
5	8	5	4
3	3	6	6
5	4	5	8
3	5	8	2
Mean = 4	Mean = 5	Mean = 6	Mean = 5
Grand Mean: 5			

AIR $(5-5)^2 + (3-5)^2 + (5-5)^2 + (3-5)^2$

$0 + 4 + 0 + 4 = 8$

BINGE $(8-5)^2 + (3-5)^2 + (4-5)^2 + (5-5)^2$

$9 + 4 + 1 + 0 = 14$

COM $(5-5)^2 + (6-5)^2 + (5-5)^2 + (8-5)^2$

$0 + 1 + 0 + 9 = 10$

DATA $(4-5)^2 + (6-5)^2 + (8-5)^2 + (2-5)^2$

$1 + 1 + 9 + 9 = 20$

Total Sum of Squares = 52

$$AIR \quad (5-4)^2 + (3-4)^2 + (5-4)^2 + (3-4)^2$$
$$1 + 1 + 1 + 1 = 4$$
$$BINGE \quad (8-5)^2 + (3-5)^2 + (4-5)^2 + (5-5)^2$$
$$9 + 4 + 1 + 0 = 14$$
$$COM \quad (5-6)^2 + (6-6)^2 + (5-6)^2 + (8-6)^2$$
$$1 + 0 + 1 + 4 = 6$$
$$DATA \quad (4-5)^2 + (6-5)^2 + (8-5)^2 + (2-5)^2$$
$$1 + 1 + 9 + 9 = 20$$

Sum of Squares Within = $SSW = 4+14+6+20=44$

SSW

$$AIR \quad (4-5)^2 = 1 \times 4 = 4$$
$$BINGE \quad (5-5)^2 = 0 \times 4 = 0$$
$$COM \quad (6-5)^2 = 1 \times 4 = 4$$
$$DATA \quad (5-5)^2 = 0 \times 4 = 0$$

Sum of Squares Between = $SSB = 4+0+4+0 = 8$

SSB

Antrix Academy of Data Science

Set Up Hypotheses

Null hypothesis
H_0 = Population means are equal.

Alternative hypothesis
H_a = Population means are *not* equal.

Rejecting H_0 indicates differences between companies.

$$F-Stat = \frac{\frac{SSB}{m-1}}{\frac{SSW}{n_t - m}}$$

Big *F*-statistic = Big difference in companies (reject H_0)

Small *F*-statistic = Not a big difference in companies (fail to reject H_0)

$$F-Stat = \frac{\frac{8}{3}}{\frac{44}{12}}$$

- m – is number of groups
- n sub t – total number of observations
- n sub t minus m is the degree of freedom

$$F-Stat = 0.727$$

Antrix Academy of Data Science

F Values for $\alpha = 0.05$

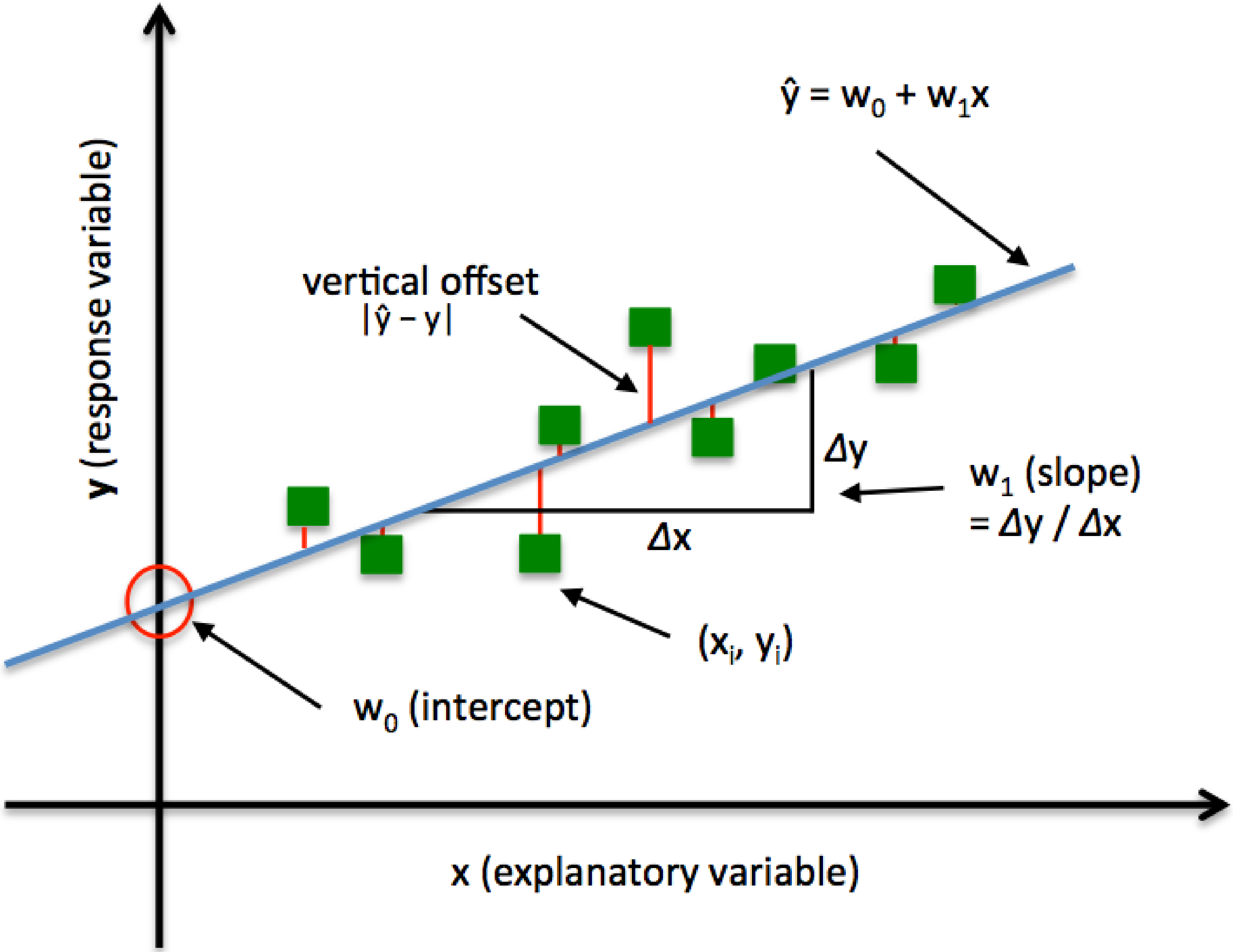
Numerator Degrees of Freedom

Denominator Degrees of Freedom df_2	df_1 1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.3	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30

- Critical F Value is 3.49,i f our F-statistic is greater than 3.49, we reject our null hypothesis. If our F-statistic is less than 3.49, we do not reject our null hypothesis.

Linear Regression

- In statistics, **linear regression** is a **linear** approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).



$\hat{y} = w_0 + w_1x$
vertical offset
$|\hat{y} - y|$
Δy
Δx
w_1 (slope)
$= \Delta y / \Delta x$
(x_i, y_i)
w_0 (intercept)
y (response variable)
x (explanatory variable)

PLAYER	HEIGHT(X)	WEIGHT(Y)	(XY)	X SQUARE	Y SQUARE
1	72	160	11520	5184	25600
2	75	180	13500	5625	32400
3	78	220	17160	6084	48400
4	77	190	14630	5929	36100
5	82	245	20090	6724	60025
SUMS	**384**	**995**	**76900**	**29546**	**202525**

$$a = \frac{n\sum(xy) - \left(\sum x\right)\left(\sum y\right)}{n\sum(x)^2 - \left(\sum x\right)^2}$$

$$a = \frac{5(76900) - (384)(995)}{5(29546)(147456)}$$

$$a = 8.832$$

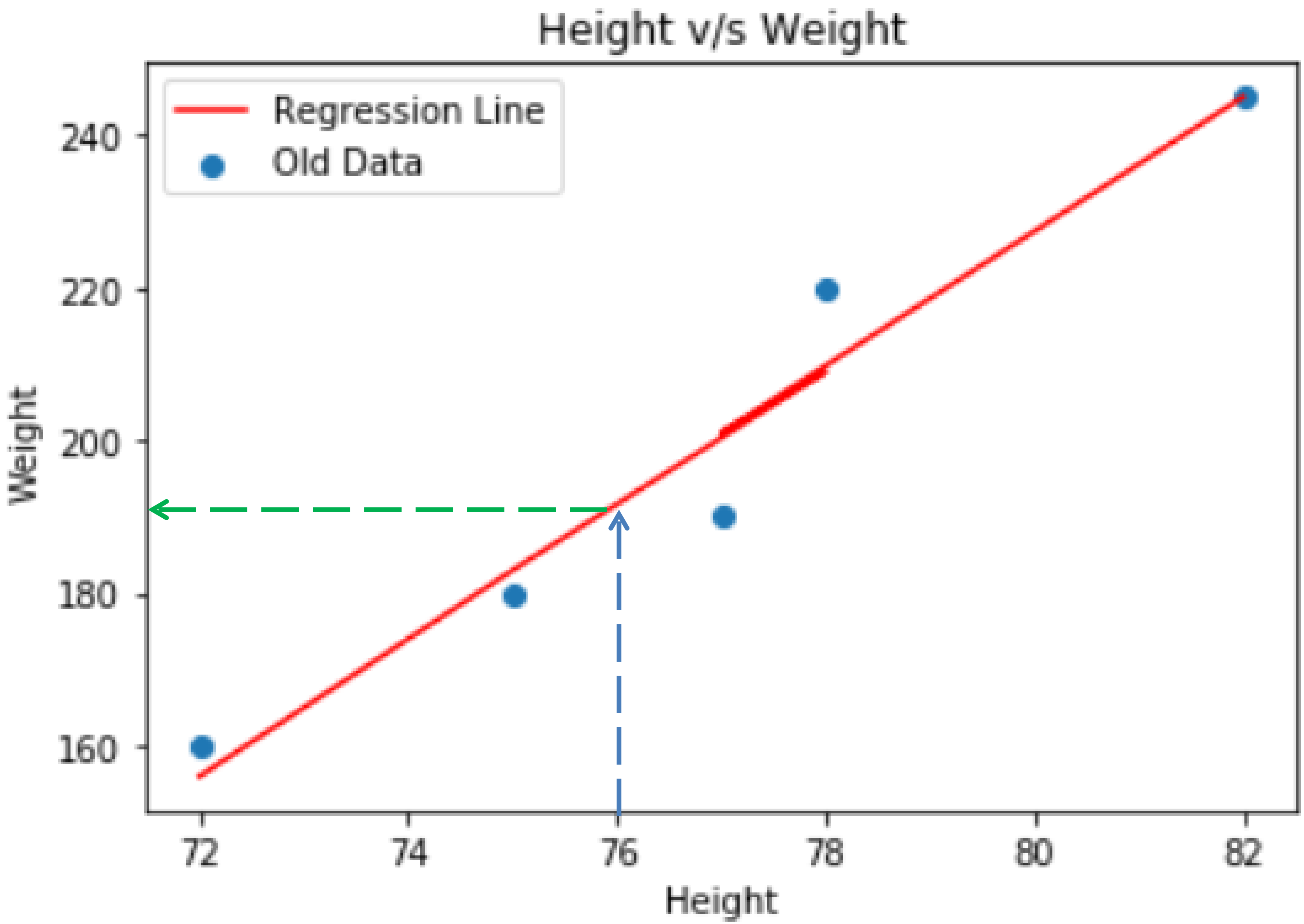
$$b = \frac{\sum y - a\sum x}{n}$$

$$b = \frac{(995) - 8.832(384)}{5}$$

$$b = -479.3$$

$$\hat{y} = 8.83x - 479.3$$

Antrix Academy of Data Science


Height v/s Weight
Regression Line
Old Data
Weight
240
220
200
180
160
72
74
76
78
80
82
Height

R-squared = SSR/SST

Sum of Squares Regression (*SSR*)

(Predicated Y – Mean Y)

Player	Height (x)	Weight (y)	Regression Squared $(\hat{y} - \bar{y})^2$
1	72	160	1809.7
2	75	180	257.6
3	78	220	109.0
4	77	190	2.6
5	82	245	2094.0
MEAN	**76.8**	**199**	**4272.8**

SSR = Sum of Squares Regression

Total Sum of Squares (*SST*)

Observed Y – Mean Y

Player	Height (x)	Weight (y)	Sum of Squares $(y - \bar{y})^2$
1	72	160	1521
2	75	180	361
3	78	220	441
4	77	190	81
5	82	245	2116
MEAN	**76.8**	**199**	**4520**

SST = Total Sum of Squares

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = \frac{4272.8}{SST}$$

$$R^2 = \frac{4272.8}{4520}$$

$$R^2 = 0.945$$

Antrix Academy of Data Science

Logistic Regression

- If response variable is not continuous value, linear regression will not be the correct model.
- When we have binary response variable, logistic regression is often used to model the data.