Variation / Dispersion

i) Range ⇒ Max(x) - Min(x) ⇒ 4

ii) Percentile ⇒ 50th Percentile

iii) Quartiles ⇒

iv) Variance ⇒

v) Standard Variation ⇒

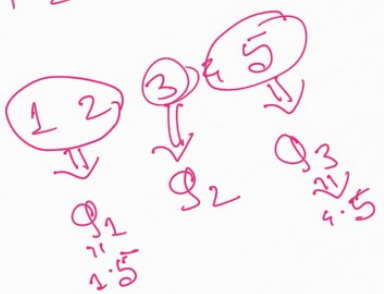
Smaller outliers < $Q_1 - 1.5 \times IQR$

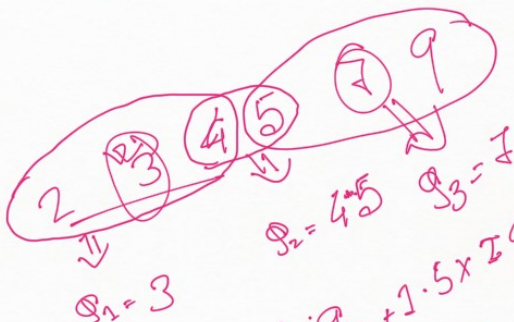
Bigger outliers > $Q_3 + 1.5 \times IQR$

IQR ⇒ Inter Quartile Range

$Q_3 - Q_1$

1 2 3 4 5





$Q_1 - 1.5 \times IQR \leq n \leq Q_3 + 1.5 \times IQR$

⇓

Allowable value in a dataset.

Dataset ⇒ X Mean ⇒ $\bar{x}$ Number of elems ⇒ n,

Absolute Deviation from A ⇒

$|x_1 - A|, |x_2 - A| \; - - - \; |x_n - A|$

Mean Absolute Deviation ⇒

$$\frac{|x_1 - A| + |x_2 - A| + \; - - - + |x_n - A|}{n} = \frac{\sum_{i=1}^{n} |x_i - A|}{n}$$

Squared Mean Absolute Deviation ⇒

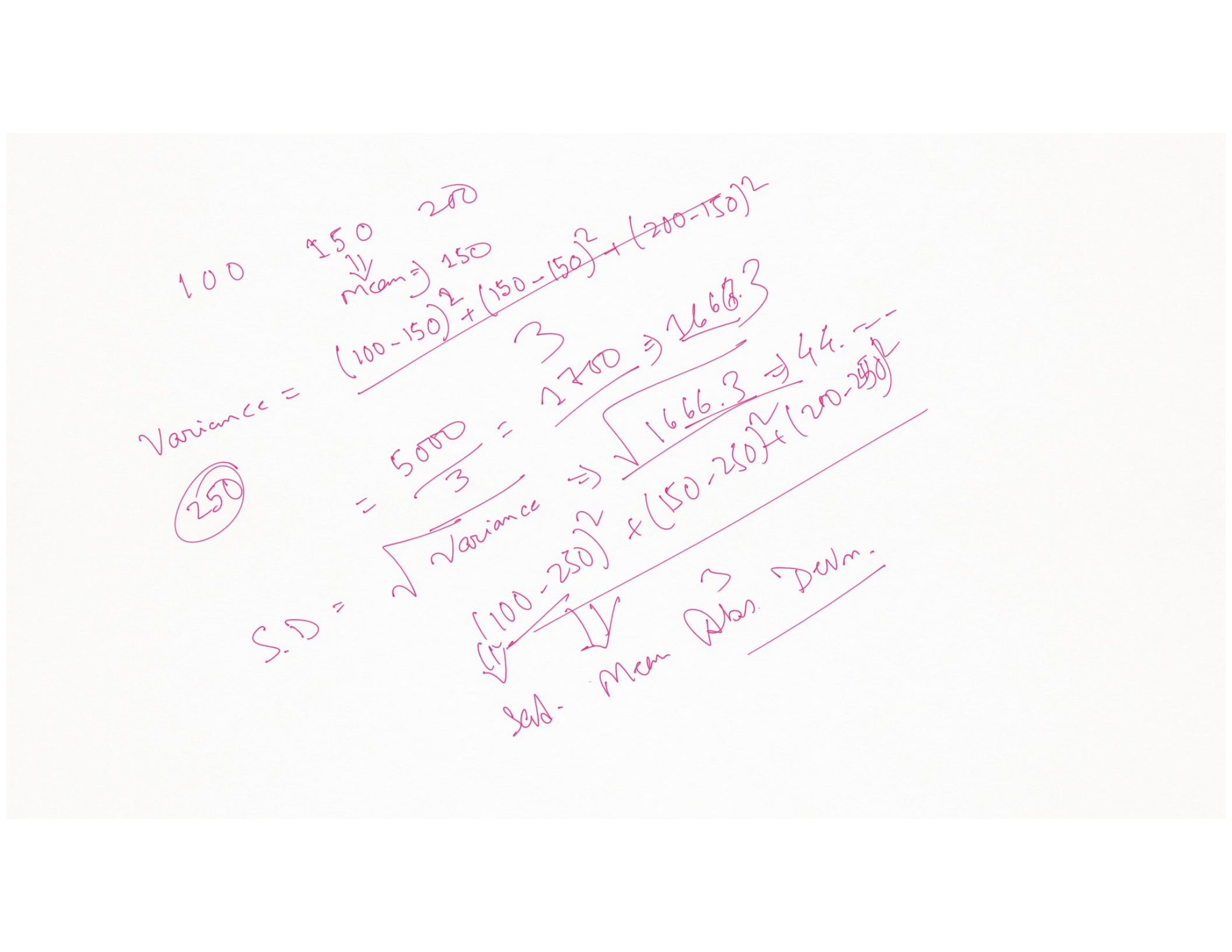
$$(x_1 - A)^2 + (x_2 - A)^2 + \; - - - (x_n - A)^2 = \frac{\sum (x_i - A)^2}{n}$$

Variance ⇒ $A \Rightarrow \bar{x} \Rightarrow \frac{\sum (x_i - \bar{x})^2}{n} \Rightarrow \sigma^2 / s^2$

Standard Deviation ⇒ $\sigma / s \Rightarrow \sqrt{\frac{(x_i - \bar{x})^2}{n}}$

1 2 3 ⇒ 5

Mean ⇒ 2


100
150
200
Mean ⇒ 150
Variance = (100-150)² + (150-150)² + (200-150)² / 3
= 5000 / 3 = 1700 ⇒ 1666.3
250
S.D = √Variance ⇒ √1666.3 ⇒ 44.---
(100-250)² + (150-250)² + (200-250)²
Std. Mean Abs Dev.

$\bar{x} = 1 \quad 2 \quad 3 \Rightarrow \sigma_x^2 \Rightarrow 2/3 \Rightarrow \sqrt{2/3}$

$\bar{y} = 3 \quad 6 \quad 9 \Rightarrow \sigma_y^2 \Rightarrow 6 \Rightarrow \sigma_y = \sqrt{6}$

$y = 3x - 2$

Covariance$(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n}$

$= \frac{(1-2)\times(3-6) + (2-2)\times(6-6) + (3-2)\times(9-6)}{3}$

$= \frac{3 + 0 + 3}{3} = 2 \Rightarrow \frac{2}{n} = 1$

$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$= (x_1 - \bar{x}) \times (y_1 - \bar{y}) + (x_2 - \bar{x}) \times (y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})$

Variance$(x) = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3}$

$= \frac{1 + 1}{3} = \frac{2}{3}$

Variance $y \Rightarrow \frac{(3-6)^2 + (6-6)^2 + (9-6)^2}{3}$

$= \frac{9 + 0 + 9}{3} = 6$

Correlation / Correlation Coefficient

Correlation Coefficient $\Rightarrow r \Rightarrow$

$$-1 \le r \le 1$$

$$r = \frac{\text{Covariance}(x,y)}{\sigma_x \cdot \sigma_y}$$

3 types correlation $\Rightarrow$ i) Negative correlation
ii) Positive "
iii) Zero "

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

$$= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{(x-\bar{x})^2 (y-\bar{y})^2}}$$

Avg Income, Crime, House Price, Number of family memb

$$\text{Covariance}(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{n}$$

$$\text{Covariance}(x,x) = \frac{\sum (x-\bar{x})(x-\bar{x})}{n}$$

$$= \frac{\sum (x-\bar{x})^2}{n} \Rightarrow \text{Variance}(x)$$

$$\Rightarrow \sigma_x$$

$$r \Rightarrow \frac{\text{Covariance}(x, y)}{\sigma_x \cdot \sigma_y}$$

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

$$= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}}$$

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

$$= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}}$$

Normal Distribution ⇒ Mean = Median = Mode = μ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\times\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\times\left(\frac{x-\mu}{\sigma}\right)^2}$$

$\frac{x-\mu}{\sigma} = z \Rightarrow$ Z-Score $\Rightarrow$ Std. Normal Variate

$x \Rightarrow$ Individual Number

$\mu \Rightarrow$ Mean

$\sigma \Rightarrow$ Std. Deviation

$$\frac{(x-\mu)}{\sigma}$$

$Q_1 = -0.675\sigma$

$Q_3 = 0.675\sigma$

$IQR \Rightarrow Q_3 - 1.5 \times IQR$

$= -0.675\sigma - 1.5(0.675\sigma + 0.675\sigma)$

$= -0.675\sigma - 2.025\sigma$

$= -2.70$

$Q_3 + 1.5 \times IQR$

$0.675\sigma + 1.5(0.675\sigma + 0.675\sigma)$

$= 0.675\sigma + 2.025\sigma$

$= 2.70$

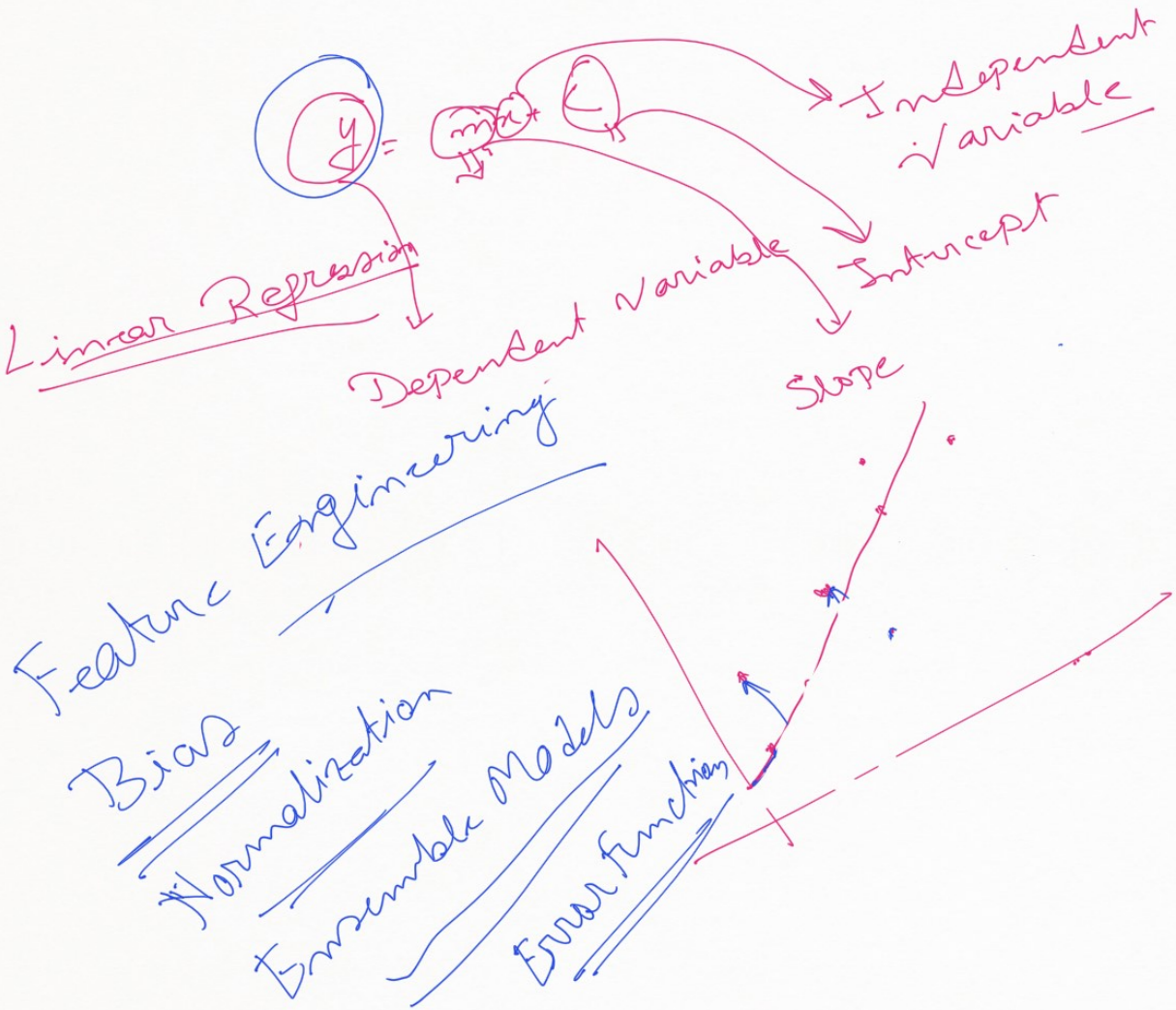
$-30 \quad 30$

99.7%

$-0.675 - 1.35$

≈ 2

≈ 2



Feature Engineering

Bias

Normalization

Ensemble Models

Error function

$y - y_p$

$n = 100 \Rightarrow y = 300$

$y = \frac{}{10}$

290