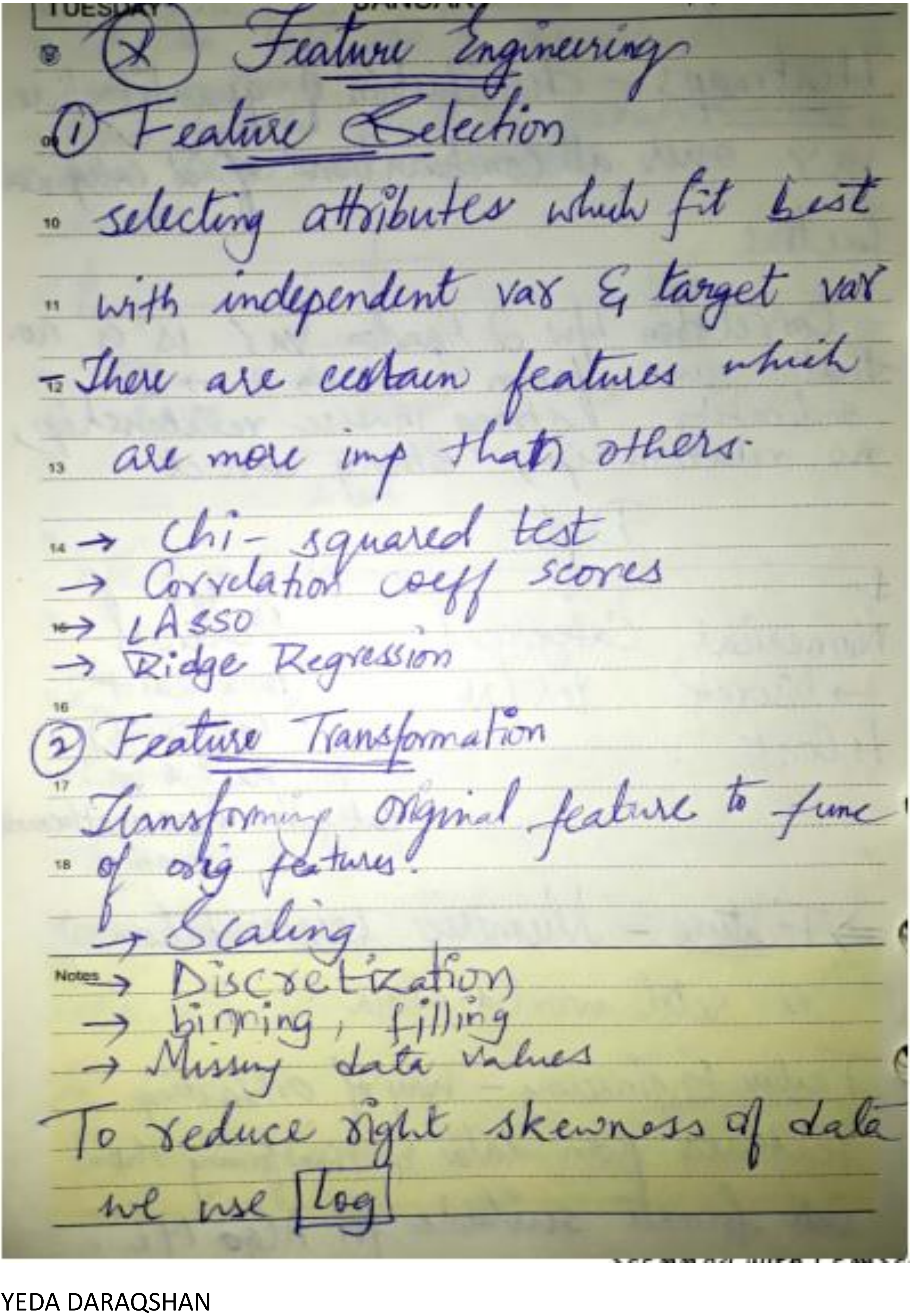


TUESDAY

(2) Feature Engineering

(1) Feature Selection

10 selecting attributes which fit best

11 with independent var & target var

12 - There are ceatain features which

13 are more imp thats others.

14 → Chi-squared test

→ Correlation coeff scores

→ LASSO

→ Ridge Regression

16

(2) Feature Transformation

17 Transforming original feature to func

18 of org features.

→ Scaling

Notes → Discretization

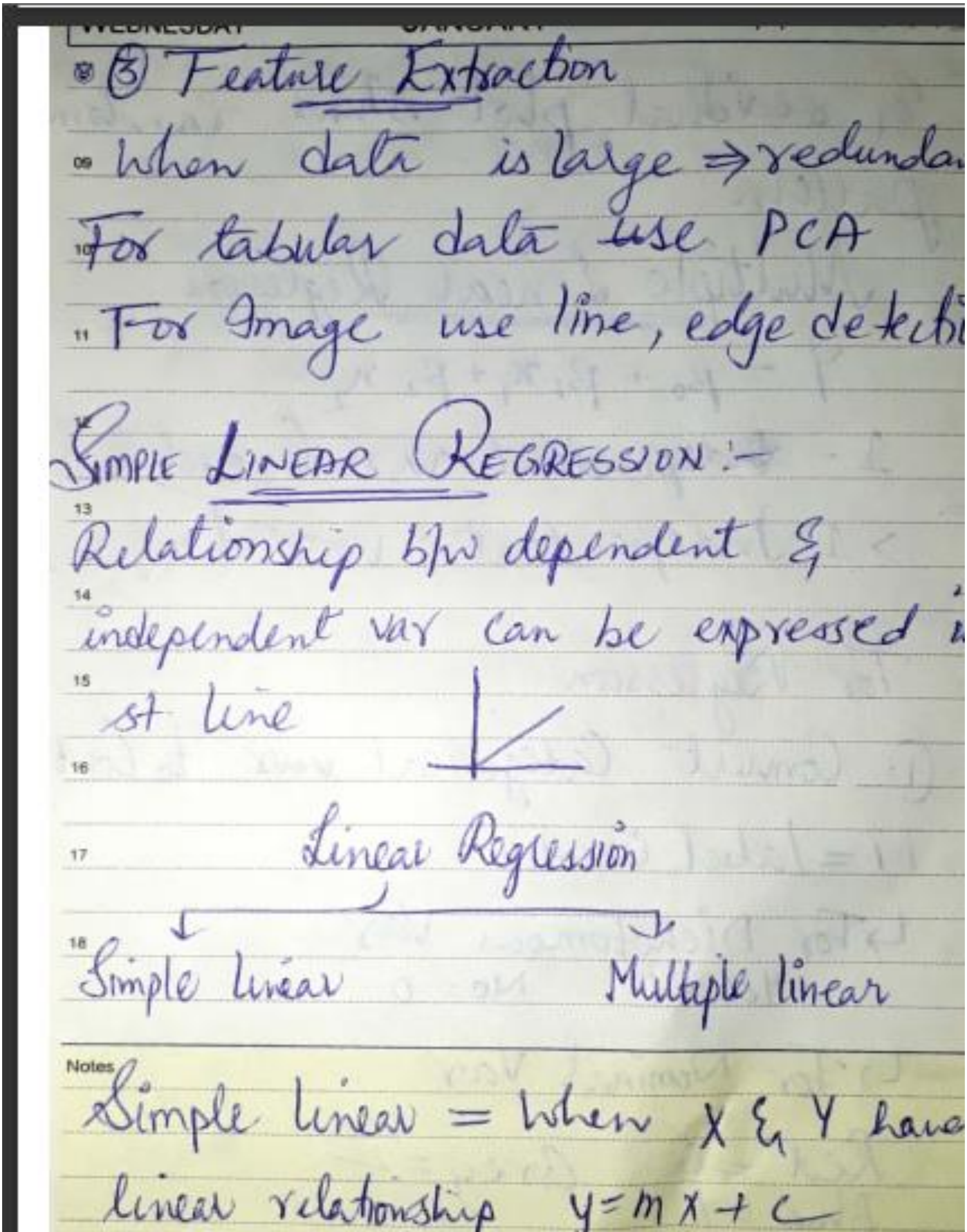
→ binning, filling

→ Missing data values

To reduce right skewness of data

we use $\boxed{Log}$

SYEDA DARAQSHAN



08 ③ Feature Extraction

09 When data is large ⇒ redunda

10 For tabular data use PCA

11 For Image use line, edge detecti

12 Simple LINEAR REGRESSION :-

13 Relationship b/w dependent &

14 independent var can be expressed i

15 st line

16

17 Linear Regression

18 Simple linear Multiple linear

Notes Simple linear = when X & Y have linear relationship $y = mx + c$

To chk this, XY Scatterplot = linea

ε_1 residual plot shows random pattern

Multiple Linear Regression

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots$

1 - dependent var (cont)

> 1 Independent var $(x_1, x_2 \ldots)$.

For Regression

(1) Convert Categorical var to Cont var

1.1 = Label Encoder

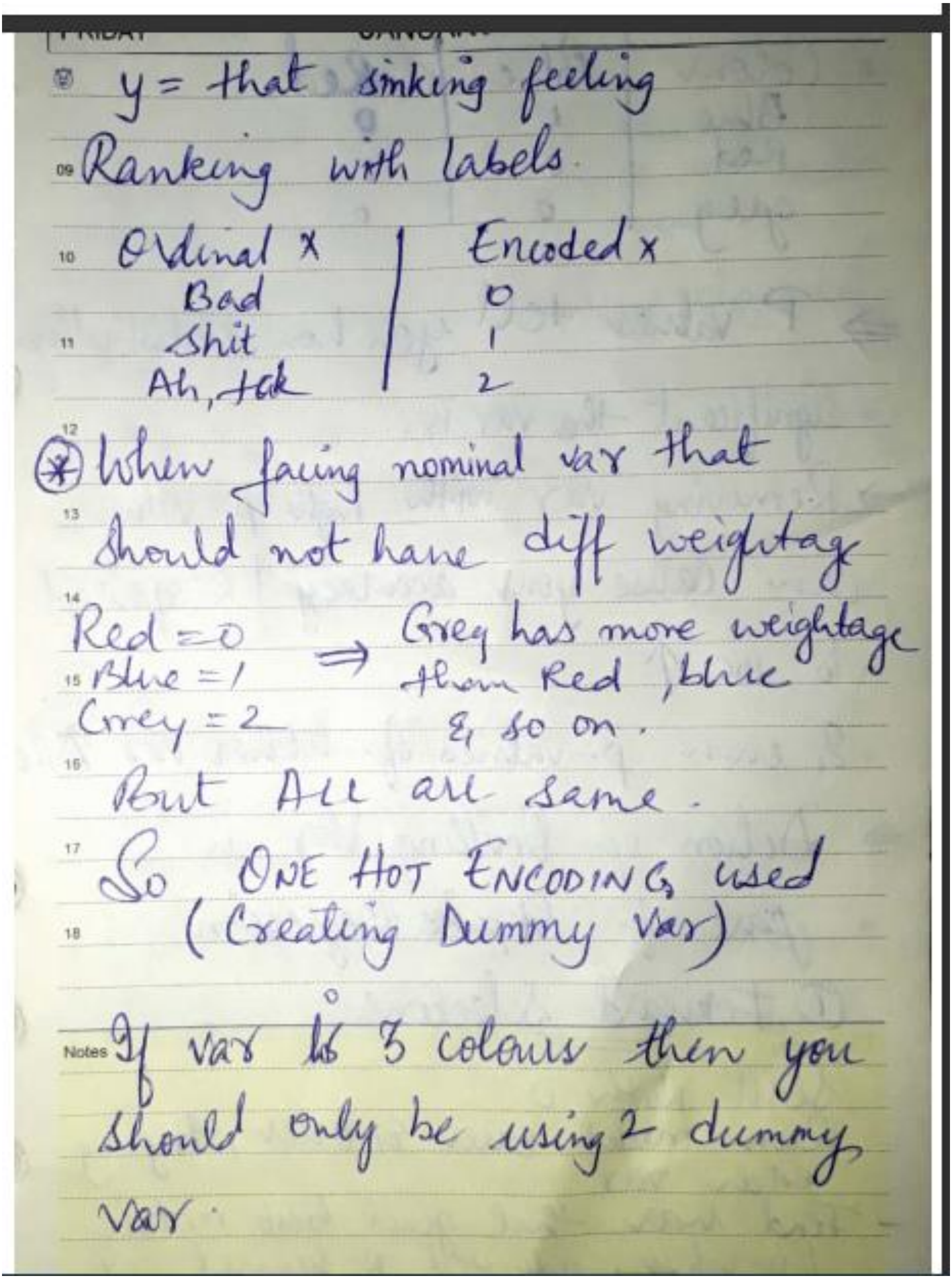
↳ For Dichotomous Var

Yes = 1 No = 0

↳ For Nominal Var

Red = 0 Grey = 2
Blue = 1

↳ For Ordinal Var



y = that sinking feeling

Ranking with labels

Ordinal x	Encoded x
Bad	0
Shit	1
Ah, fck	2

⊛ When facing nominal var that should not have diff weightag

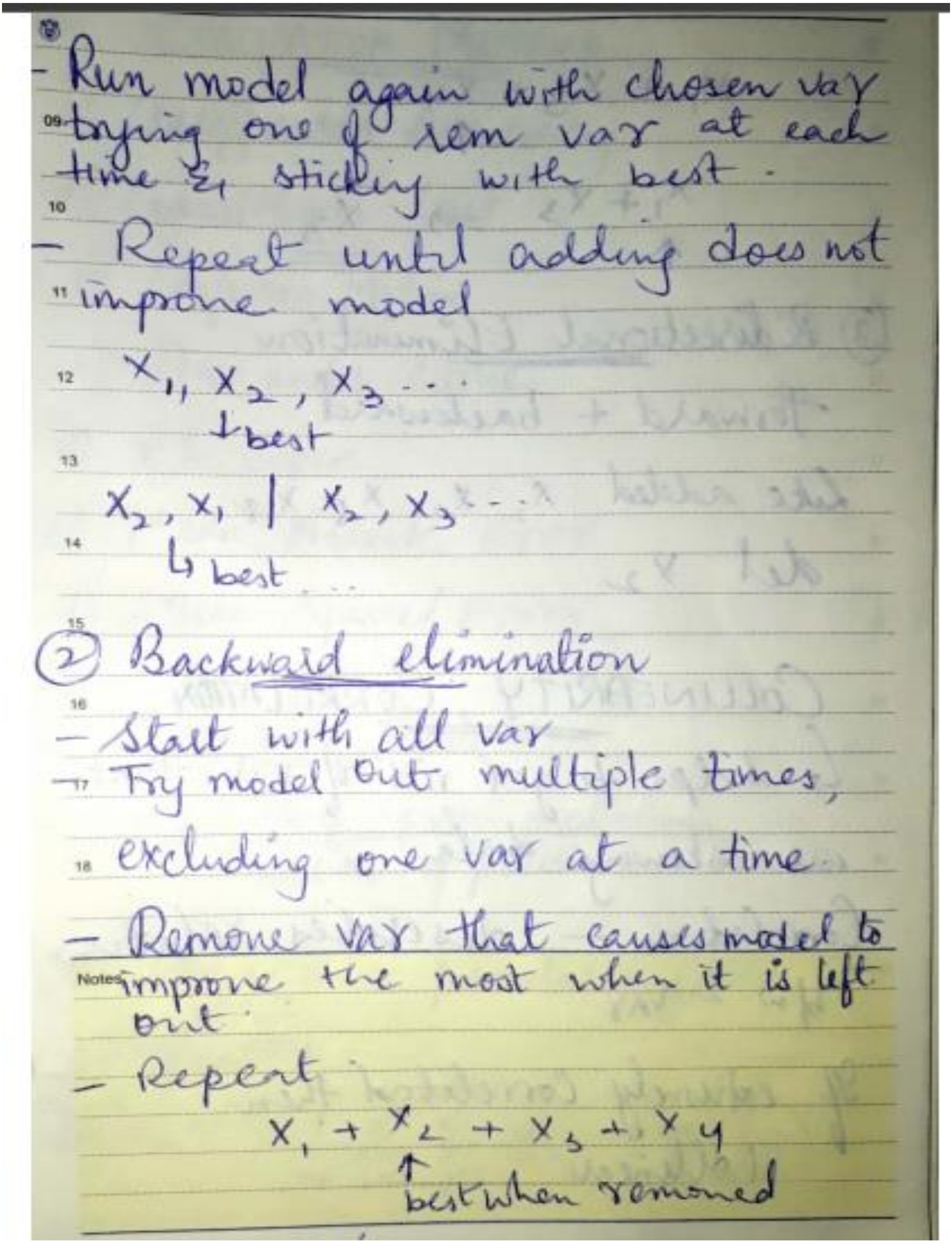
Red = 0
Blue = 1
Grey = 2

⇒ Grey has more weightage than Red, blue & so on.

But ALL are same.

So ONE HOT ENCODING used (Creating Dummy Var)

If var is 3 colours then you should only be using 2 dummy var.



- Run model again with chosen var trying one of rem var at each time & sticking with best.

- Repeat until adding does not improve model

$x_1, x_2, x_3 \dots$

$\downarrow$ best

$x_2, x_1 \mid x_2, x_3 \dots$

$\hookrightarrow$ best ...

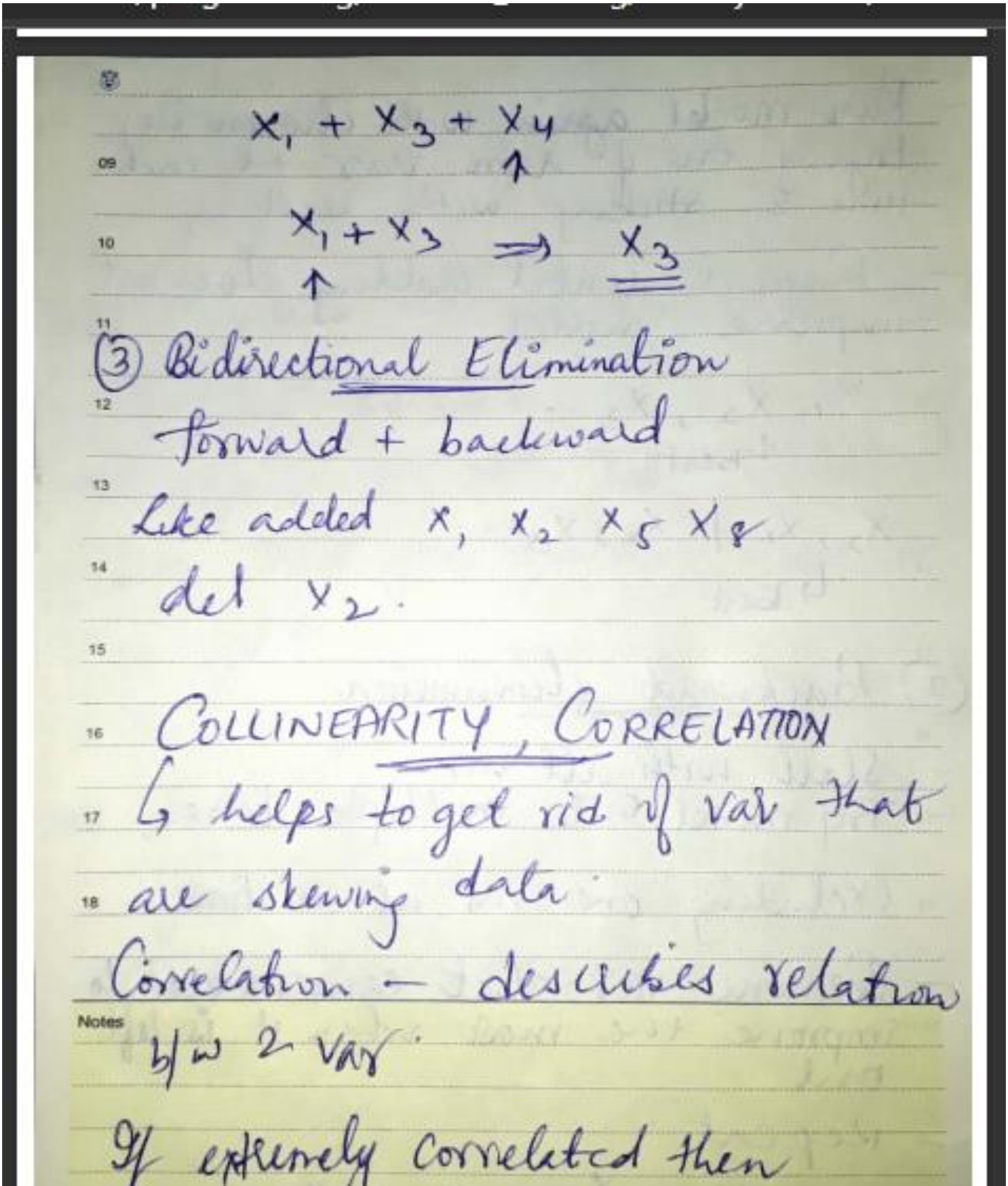
② Backward elimination

- Start with all var
- Try model out multiple times, excluding one var at a time.
- Remove var that causes model to improve the most when it is left out.
- Repeat

$$x_1 + x_2 + x_3 + x_4$$

$\uparrow$ best when removed

SYEDA DARAQSHAN



$$X_1 + X_3 + \underset{\uparrow}{X_4}$$

$$\underset{\uparrow}{X_1} + X_3 \Rightarrow \underline{\underline{X_3}}$$

(3) Bidirectional Elimination

forward + backward

Like added x_1 x_2 x_5 x_8

del x_2.

COLLINEARITY, CORRELATION

↳ helps to get rid of var that are skewing data

Correlation – describes relation b/w 2 var.

If extremely correlated then collinear