When tomorrow is the last date of assignment submission

UNIVERSITY STUDENT.org

upGrad

#LifeKoKaroLift

Data Science Certification Program

upGrad

Course : Machine Learning

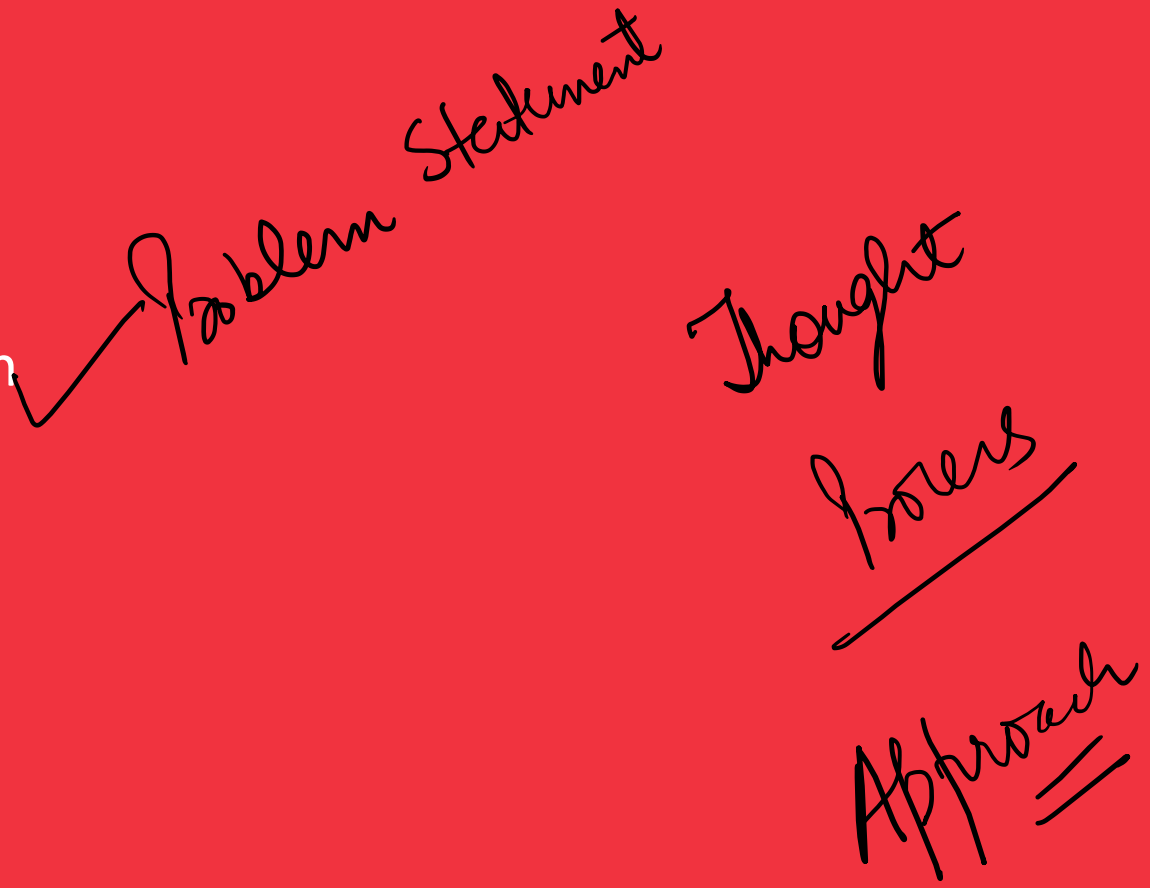
Lecture On : Linear Regression Assignment

Instructor : Shivam Garg

upGrad

Today's Agenda

1. Assignment Walkthrough
2. Step by Step Approach
3. Doubt Session (QnA)



Linear Regression:

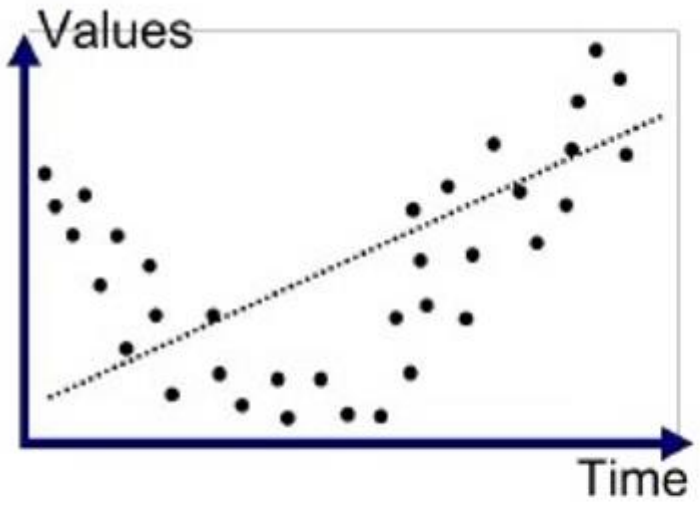
upGrad

Quick Revision

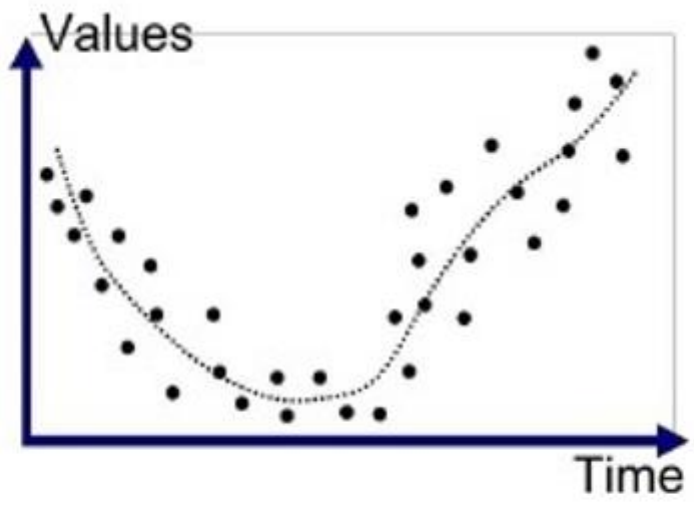
Years of Experience	Salary in 1000$
2	15
3	28
5	42
13	64
8	50
16	90
11	58
1	8
9	54

Linear Regression:

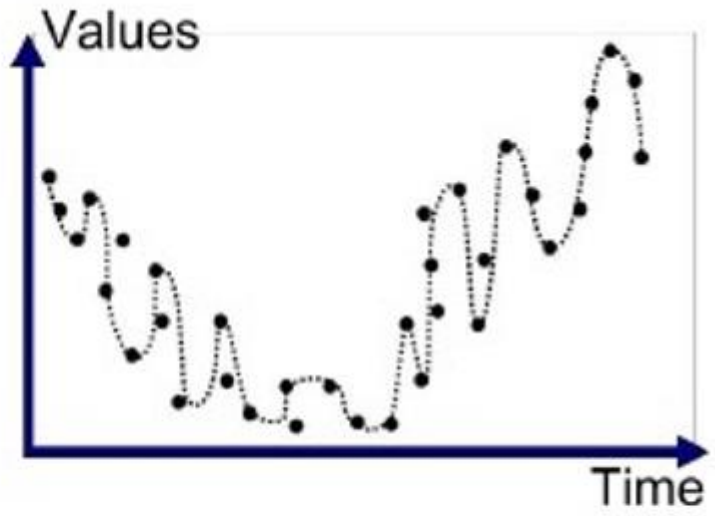
upGrad



Underfitted



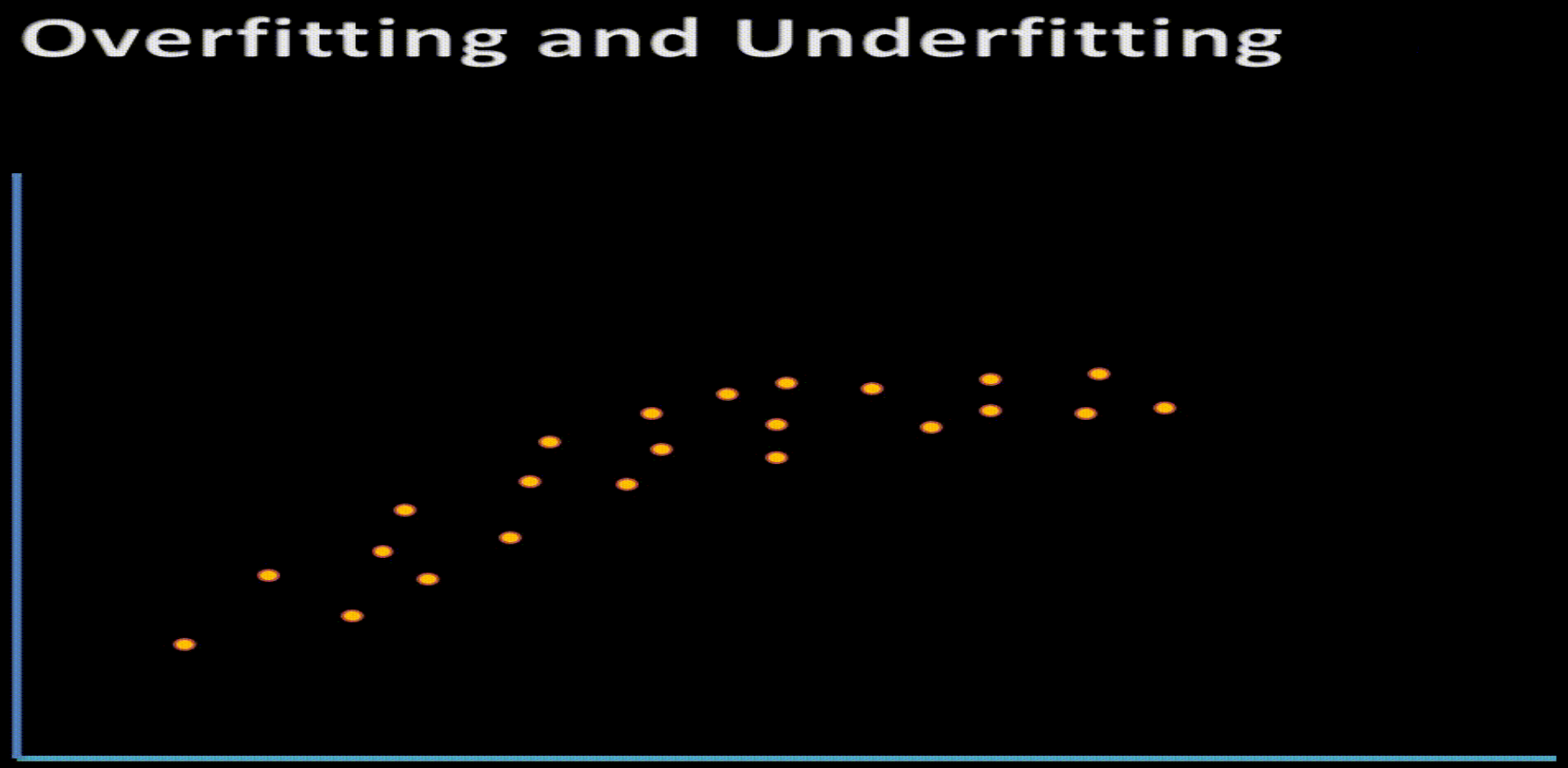
Good Fit/Robust



Overfitted

Linear Regression:

upGrad



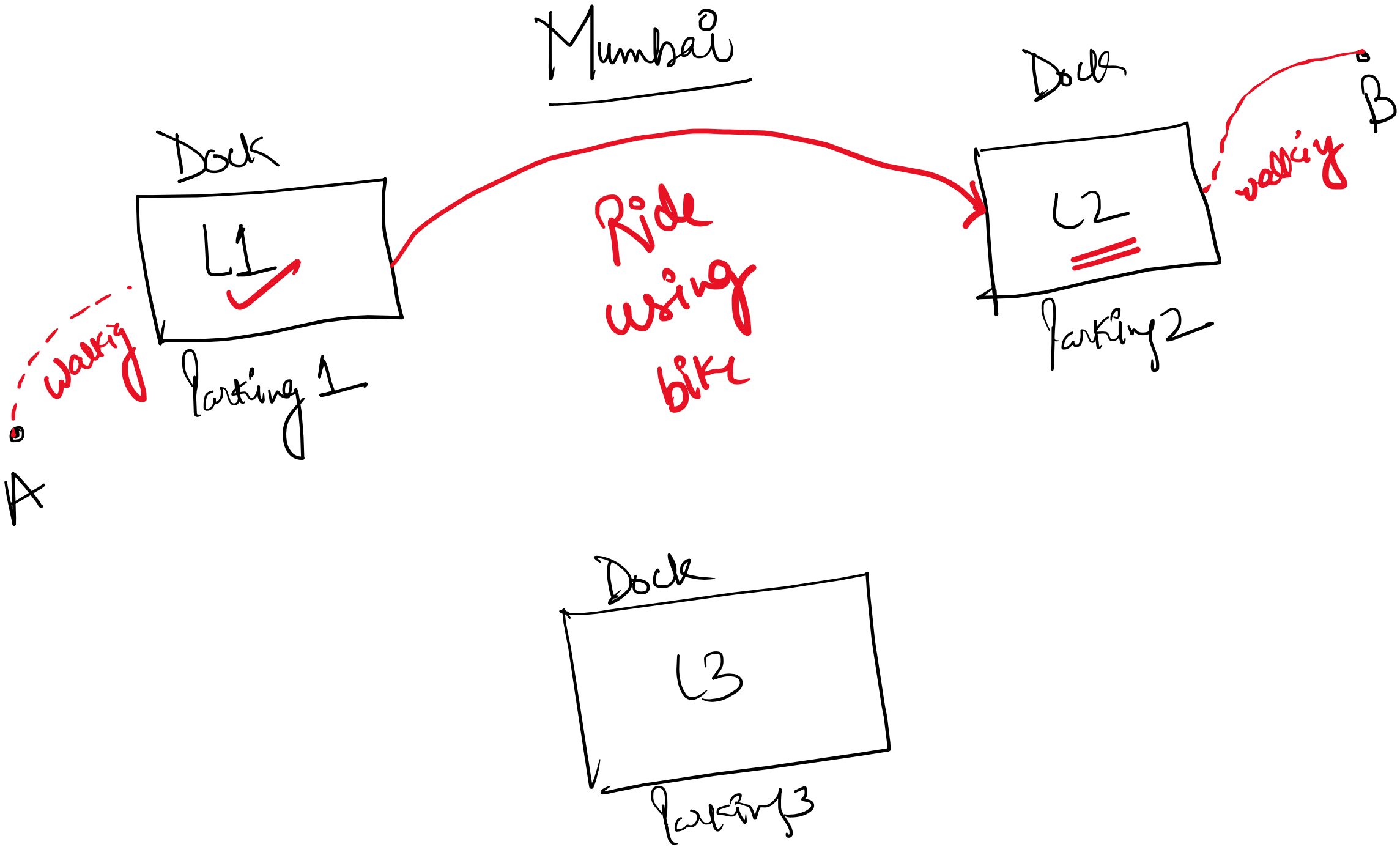
Linear Regression:

upGrad

Assignment Problem Statement

- A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short-term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system.

- A US bike-sharing provider BoomBikes has recently suffered considerable dips in their revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain in the current market scenario. So, it has decided to come up with a mindful business plan to be able to accelerate its revenue as soon as the ongoing lockdown comes to an end, and the economy restores to a healthy state.


Mumbai
Dock
L1
Parking 1
Dock
L2
Parking 2
Dock
L3
Parking 3
Walking
Ride using bike
Walking
A
B

Linear Regression:

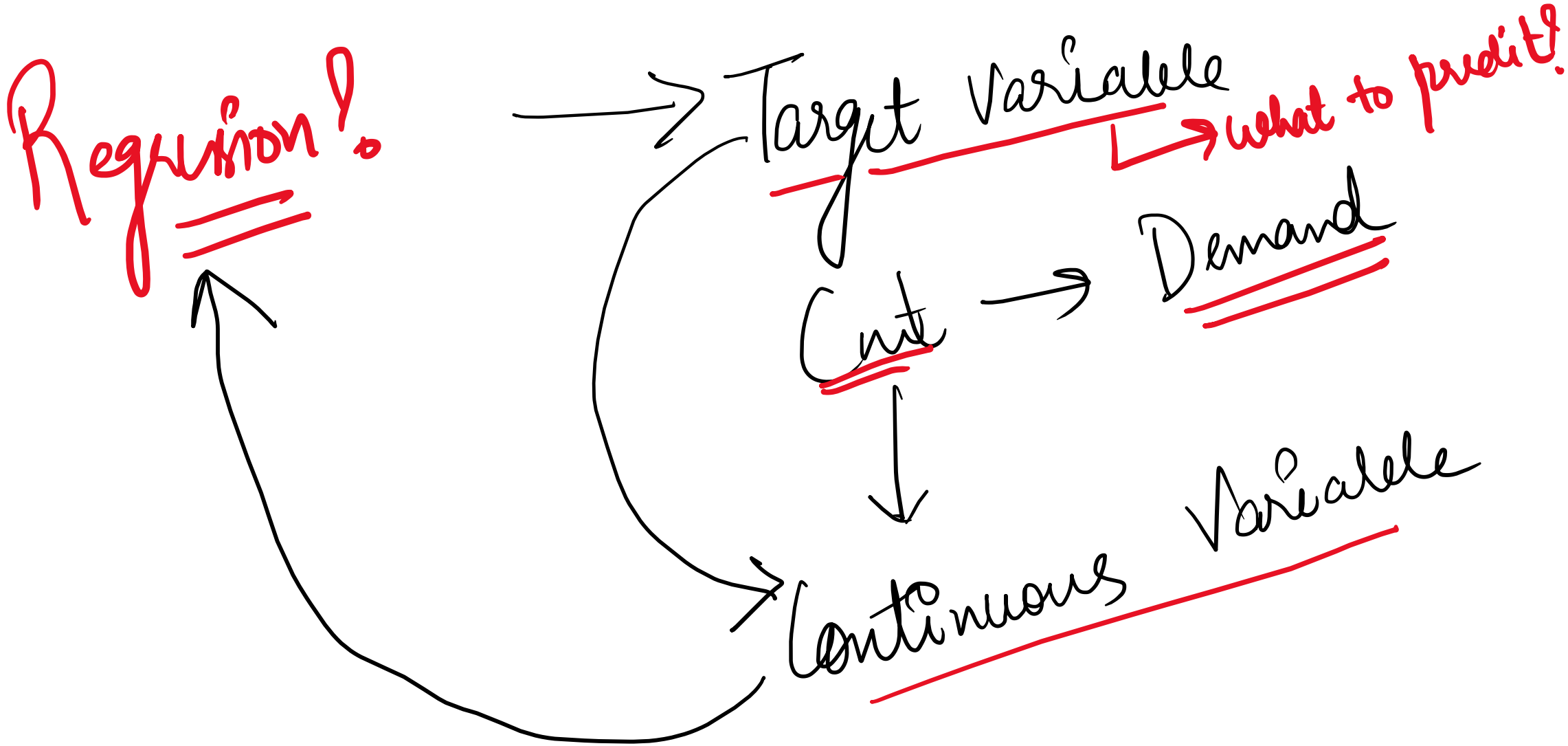
upGrad

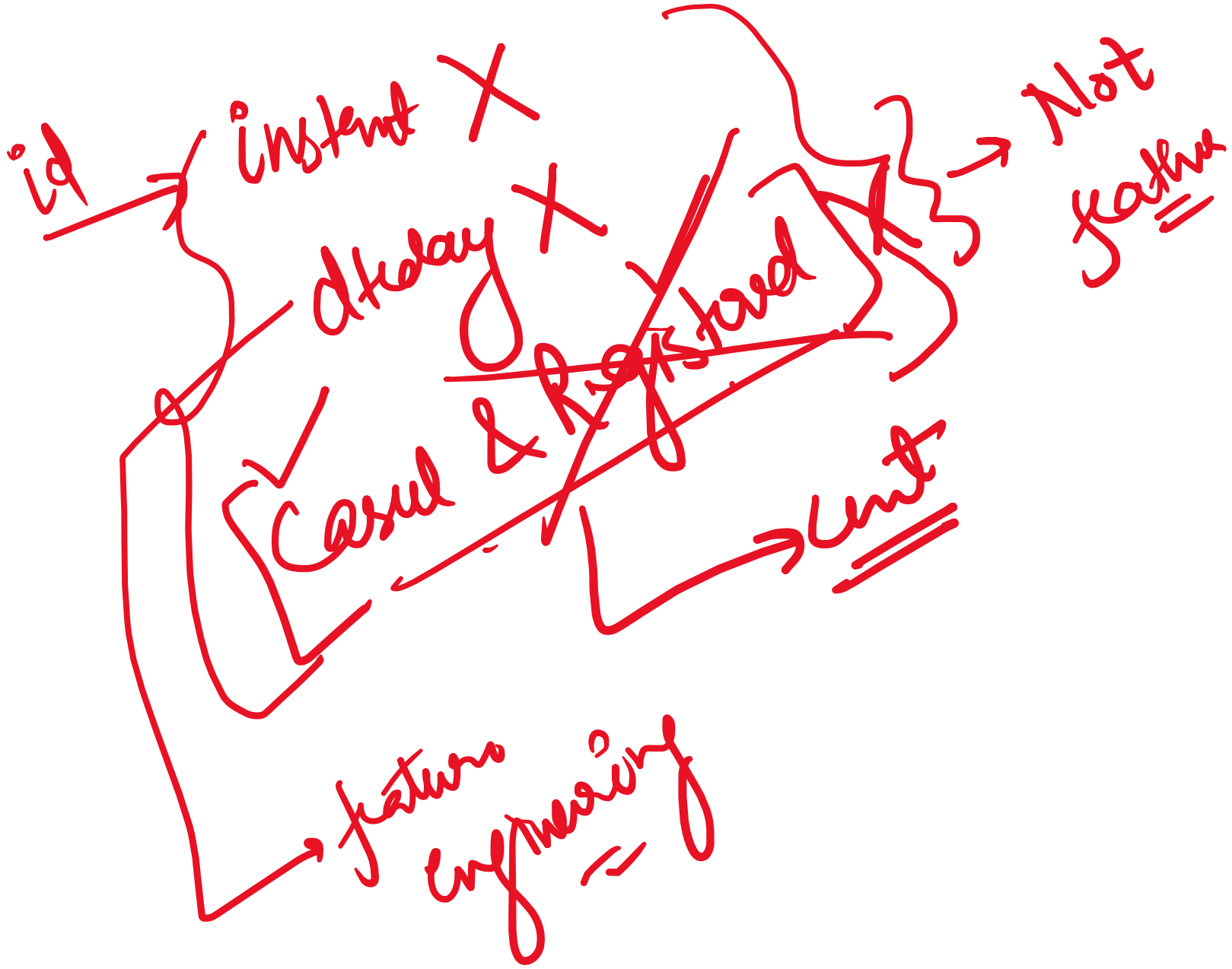
- They have contracted a consulting company to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know:

1. Which variables are significant in predicting the demand for shared bikes.
2. How well those variables describe the bike demands

What you need to do?

- Create a linear model that describe the effect of various features on price.
- The model should be interpretable so that the management can understand it.


Regression!
Target Variable
what to predict!
Cnt
Demand
Continuous Variable


id
instant X
delay X
casual & Registered
Not feature
count
feature engineering

① Understand the problem Statement & understand the data dictionary.

② Load the data

→ pd.read_csv( )

③ Sanity check

(i) Shape of the data

(ii) No of columns

(iii) Missing values

(iv) data types etc etc

④ Data preprocessing

↳ data driven

↳ Encoding

Weather

1

2

3

4

Weather

1 → summer

2 → winter

3 → Rainy

4 →

⑤ Dummy Creation → Only categorical variables.

Unordered	Weathersit
1	1
2	2
3	3
4	4
5	2
6	3
1	1
8	2

No of unique values = 4 = n (drop_first = True)

No of dummy = 3

3 dummies

(n-1)

	W-1	W-2	W-3	W-4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	0	1	0	0
6	0	0	0	0
7	1	0	0	0
8	0	0	0	0

⑥ EDA

Exploratory Data Analysis (Univariate + Bivariate)

⑦ Train-Test Split

Data

Stratify

Random Split

Train data → Train my model

70:30, 75:25, 80:20

↳ [most preffered]

Test data

↳ evaluate my

from sklearn.preprocessing import train_test_split

random_state = 40

(70, 30)

model

Stratify = ——

Notebook gpu off

random

Train ✓

Test ✓

Steps

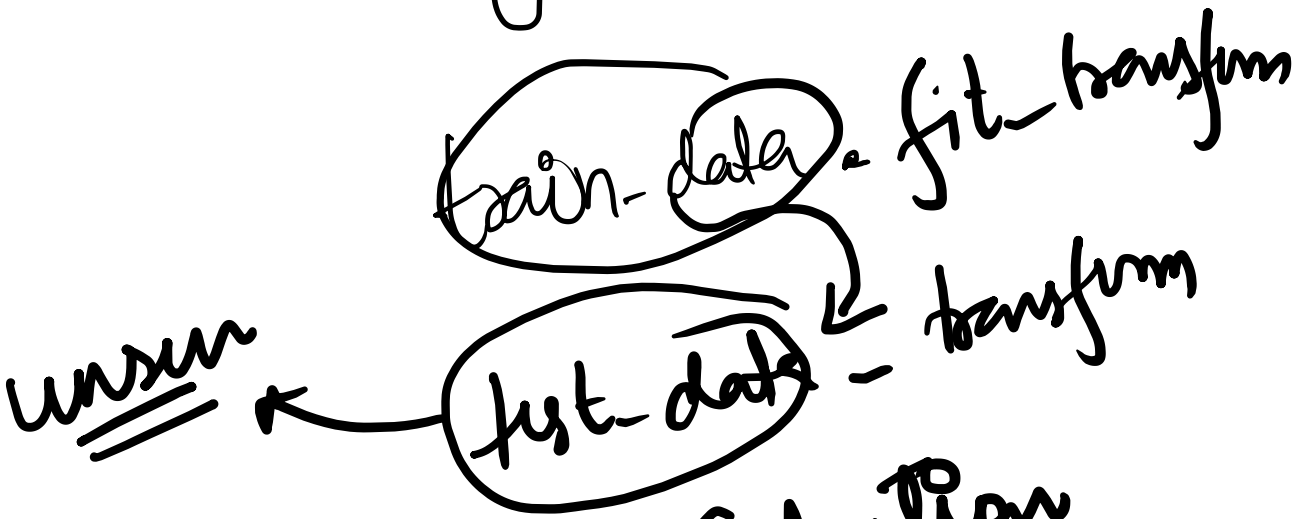
modelling part

Results

Random state = 0

Reproducibility → Randomness.

⑧ Scaling → MinMax Scaler / Standard Scaling



⑨ Feature Selection

fit

transform

evaluating $\bar{x}, \sigma$

$\frac{20-25}{30} = \frac{5}{30} = \frac{1}{6}$

$\frac{x-\bar{x}}{\sigma}$

$\frac{x-\bar{x}}{\sigma}$

fit data

30 → 5

$\frac{30-25}{30} = \frac{5}{30}$ 20 → 1/6 Scaling

$\frac{30}{1} = 30$

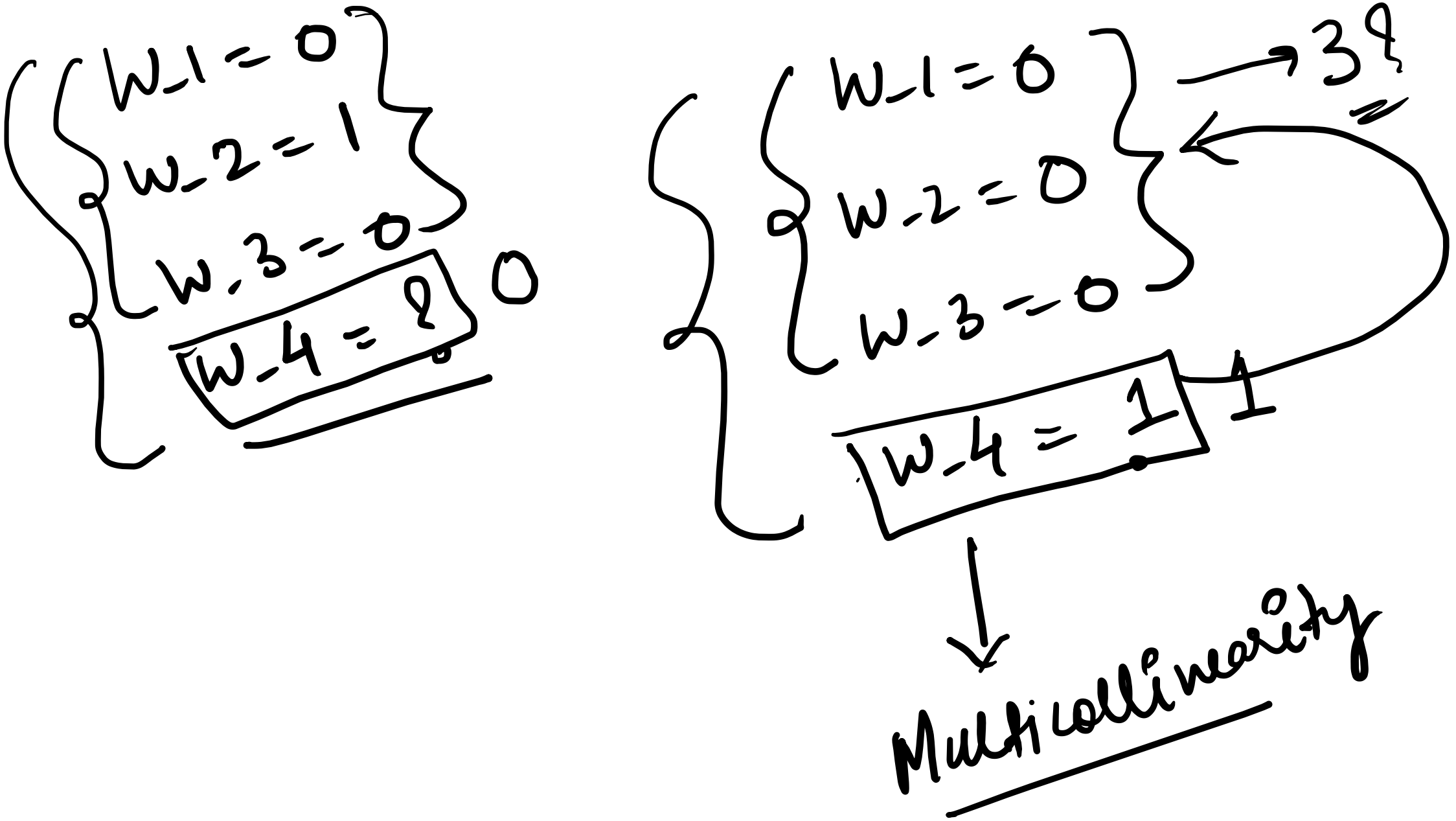
→ 30

$30-30 = 0$

Product

↳ you will get one

list


W_1 = 0
W_2 = 1
W_3 = 0
W_4 = 0
0
W_1 = 0
W_2 = 0
W_3 = 0
W_4 = 1
1
Multicollinearity

⑨ Feature Selection

(i) RFE → Automatic way of feature Selection

(ii) VIF → Manual method
- → p-value / VIF

(iii) Hybrid (RFE + VIF)

→ 33 features

→ RFE (15) variables → 15

→ keep 15

→ VIF → 15

VIF > 5

p-value > 0.05

⑩ Modelling & Evaluate

RFE → 15 → feature set

→ Modelling

OLS.summary( )

→ p-value

< 0.05

→ drop that

→ ok

Drop one feature

✓ X1

~~X2~~

X3

0.06 } altogether → same step

0.1 }

0.05

0.04 ✓

X1

0.05 ✓

X3

X p-value > 0.05 ∃ exception

drop

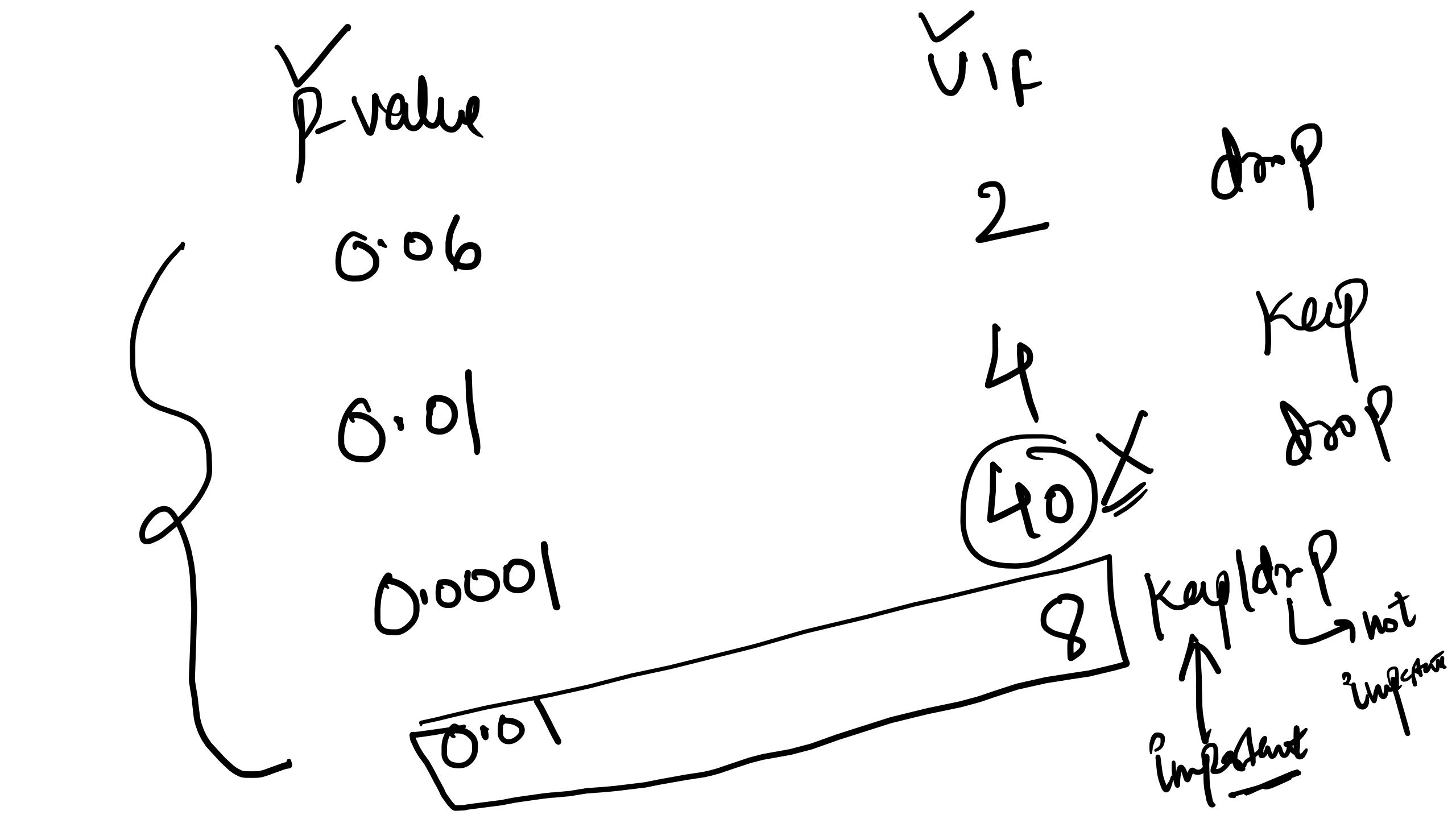
VIF > 5

Yo X

↳ Variable is very important from business respective you still can keep it

7

8


p-value
0.06
0.01
0.0001
VIF
2
4
40
8
0.01
drop
keep
drop
keep/drop
not important
important

Adj R^2

① Residual Analysis

(i) Multicollinearity

(ii) Plot y_{pred} vs y_{actual}

(iii) Plot the error term.

VIF

↓

Multicollinearity

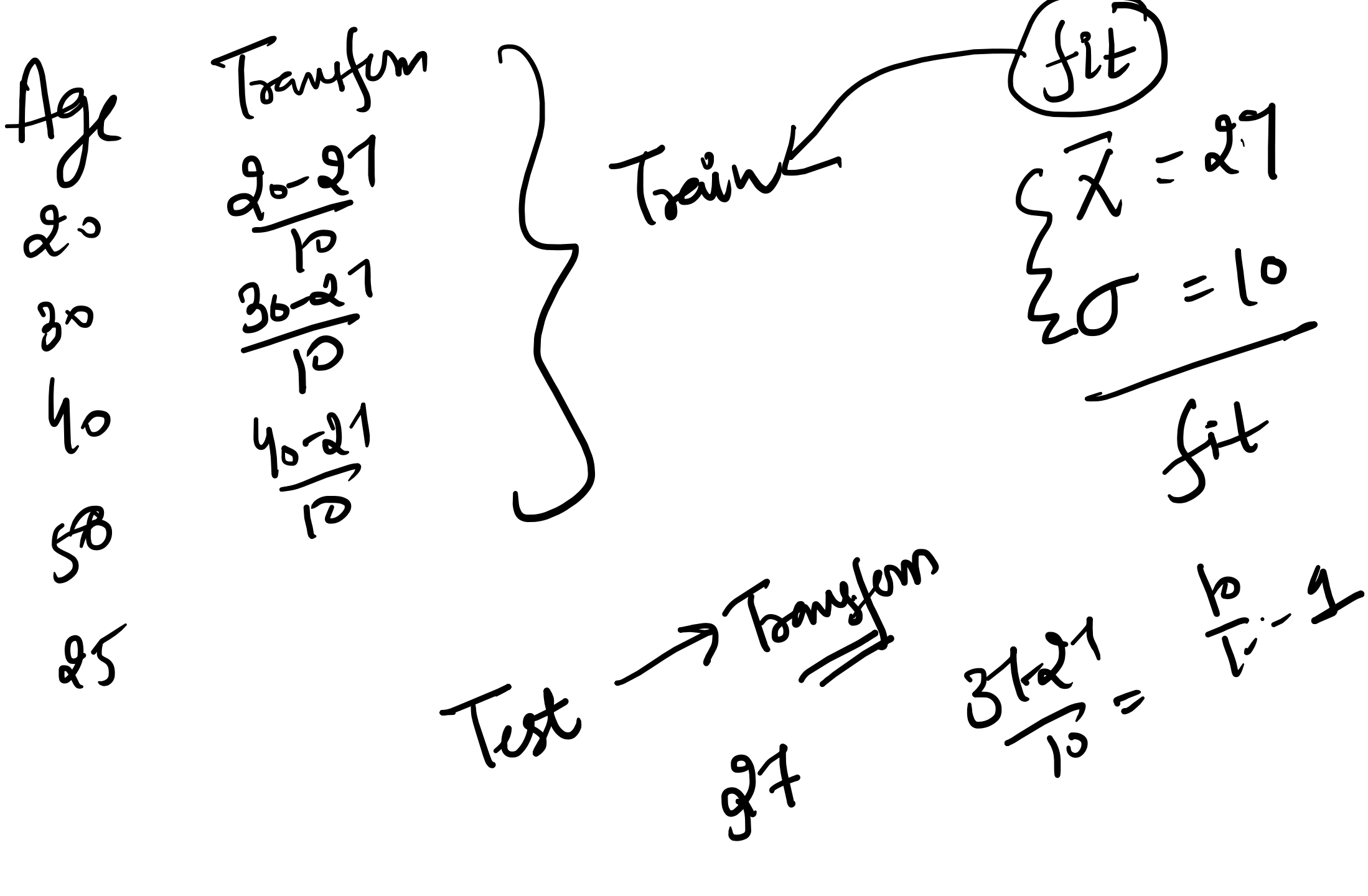
↓

two independent features are related with each other

$$= \frac{1}{1-R^2}$$

R^2 = +ve

$$Adj\ R^2 = 1 - \frac{(1-R^2)(N-1)}{(N-P-1)}$$


Age
Transform
20
30
40
50
25
20-27/10
36-27/10
40-27/10
Train
fit
$\bar{x} = 27$
$\sigma = 10$
fit
Test
Transform
27
31-27/10 =
1/10 - 1

Test data ⟶ 1 data point

(1) Age

⟶ 30

$\bar{X} = 30$

30 − 30 → train data

$\frac{0}{\sigma} = 0$

(2) 35

35 − 35 → train data

$\frac{0}{\sigma} = 0$

Final No of features 6–10

Adjusted R^2 ≈ 80%

[78 - 82]

Linear Regression:

upGrad

Data Preparation:

- Load the data and understand it using dictionary provided.
- Convert the columns to proper data types.
- Create dummies for categorical variables.

Model Building

- Divide the data to train and test.
- Perform scaling.
- Divide the data into X and y.
- Perform Linear Regression.
- Use mixed approach if you want.

Linear Regression:

upGrad

How to go about selecting features for a good model?

- RFE
- Manual
- Mixed

upGrad

Assignment Steps

Model Evaluation

- Check the various assumptions.
- Check the Adjusted R-Square for both test and train data.
- Report the final model.

upGrad

Assignment-Subjective

Steps to answer subjective part

- Answer all the questions.
- You can write the answer using any software but submit the file in PDF format
- You can use images and plots to support your answer.
- Make sure the question is answered with sufficient number of word: No limit
- Please don't copy for any online available literature.

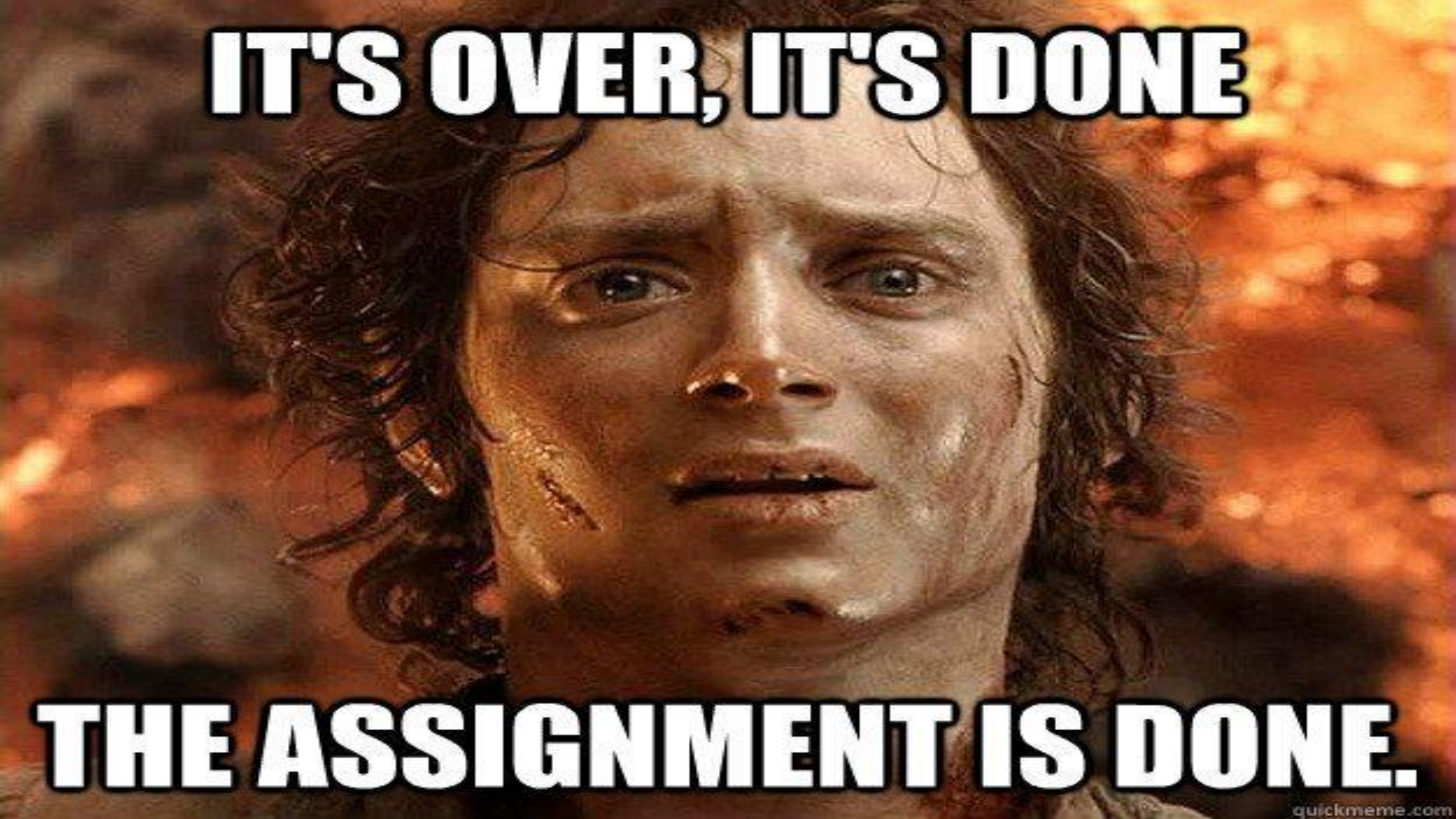
Linear Regression:

upGrad

Assignment-Endnote

What to keep in mind

- Add comments after every cell of code. So that we can understand your approach and method.
- Describe the results.
- For subjective answers, use DOC and type on it, if you wish to add images you can. But convert it to PDF before submitting.
- Create only one Jupyter notebook.
- Submit one zip file with the code and the PDF.
- Use StackOverflow for dealing with syntax errors. Rather than being stuck at one place or waiting for someone to resolve your doubts, take action and use the resources available on the internet to save time.
- Post on the discussion forums for resolving any doubts you have
- Finally, write code manually instead of copy-pasting from the in-content notebooks provided. Builds a habit of writing code. It's okay to look and write, but don't just copy-paste under any circumstance. Because of just copy-pasting, a lot of our students have faced difficulties in the past when they had to write some code on their interview.


IT'S OVER, IT'S DONE
THE ASSIGNMENT IS DONE.
quickmeme.com