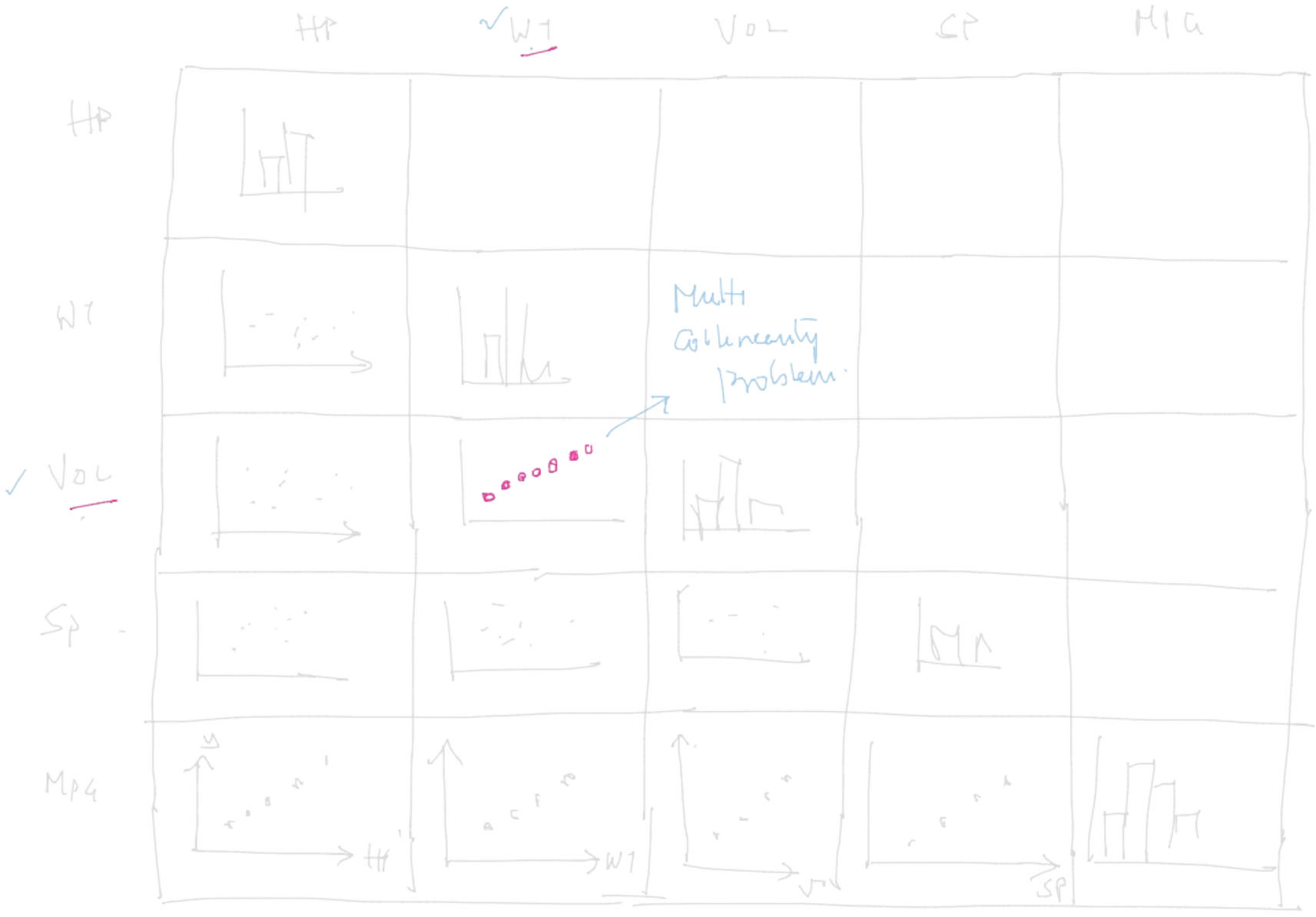

HP
WT
VOL
CP
MPG
HP
WT
✓ VOL
SP
MP4
Multi Collinearity problem.
HP
WT
SP

→ Correlation

$$y = \beta_0 + \beta_1 WT + \beta_2 VOL + \beta_3 SP + \beta_4 HP$$

$\beta_1 = 0$; $\beta_2 = 0$;

"WT affects MPG"

WT - | MPG

$\Rightarrow R^2 = 0.277$

"VOL affects MPG"

VOL | MPG

$R^2 = 0.28$

WT VOL | MPG

Expected $R^2 = 50\%$

$R^2 = .28$

Worse

Variance Inflation Factor (VIF).

VIF of SP → How much Correlation SP has with other X-Values → Model.

Ignore y (MPG).

$\Rightarrow R^2 \uparrow$ SP Strongly Correlated

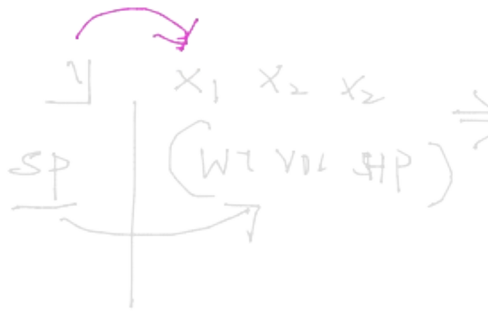
$VIF \Rightarrow \frac{1}{1-R^2}$

$R^2 \downarrow$ No Correlation between SP and other X-Values

$R^2 \rightarrow$ % Variance ✓

$R^2 \uparrow$, Hpn x, y

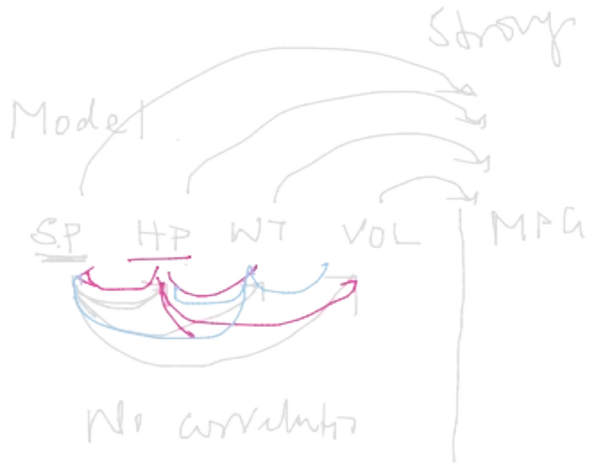
$R^2 \downarrow$



$\Rightarrow R^2$

$R^2 \downarrow$

R^2



R^2 less — SP - good feature.

VIF less

$R^2 = 0$ — HP is good feature

$R^2 = 0$ — WT is good

$R^2 = 0$ — VOL is good predictor

$$\uparrow VIF = \frac{1}{(1 - R^2\uparrow)\downarrow}$$

Large VIF Bad Factor

(1) Decided The best Feature

SP	VOL	HP	MPG

✓ 1. Strong correlation betn x & y

✓ 2. No correlation betn x-values.

→ removed WT, so now there is no correlation betn x-values.

(2)

Error

Constant Variance

Mean = 0

+ve

-ve

+2L

-2L

Homoscedasticity

3 Residuals.

✓ Mean = 0 ✓

✓ Constant Variance

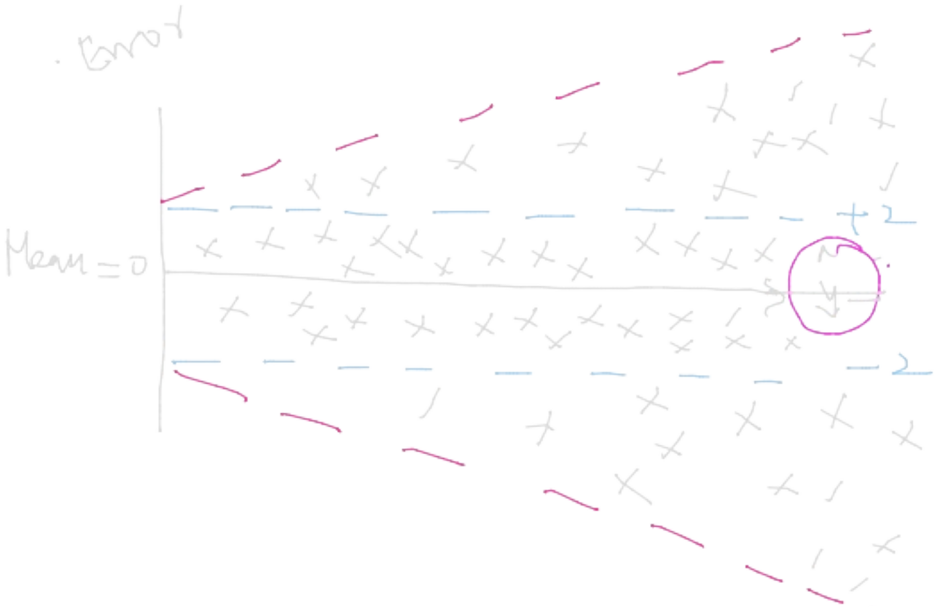
✓ Normally distributed ✓

No auto correlation ✓

y	$\hat{y}$ → Error
50	45+5 +ve
to	43+5 +ve
70	58+1 +ve

Mean = +ve

No Constant Variance



Funnel → Heteroscedasticity → Error dist → Varying

Random distribution

3σ → 3σ Lacy

3σ → 2L

±2L → γ

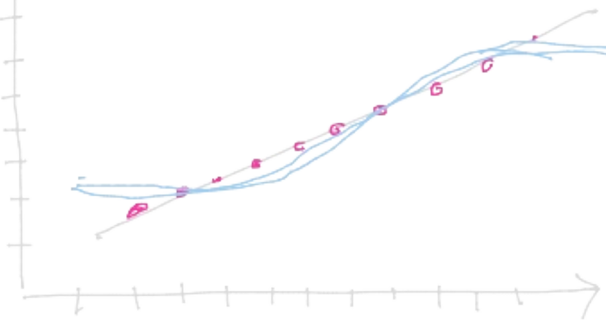
Homoscedasticity → Error distribution → uniform

Quantile – Quantile Plot

Q – Q plot

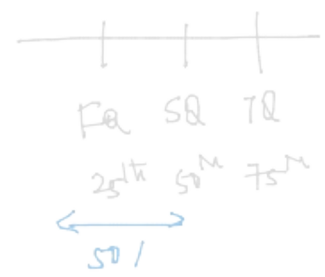
— check if your dataset follows a particular distribution

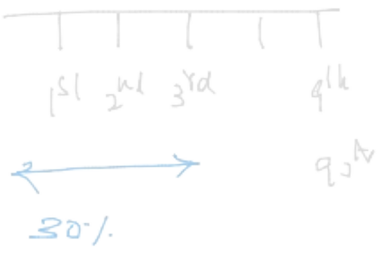
Error quantiles



QuaNtile

- Percentiles
- QuaRtile
  - Fq SQ TQ
  - 25th 50th 75th
- Deciles





Error	Normal dist
—	—
—	—
—	—
—	—
—	—
—	—
—	—
—	—

Percentile Errors	Percentile of ND
1st —	— 1st
2nd —	— 2nd
3rd —	— 3rd
—	—
—	—

Create a Normal distribution

⇒ Quantile Quantile

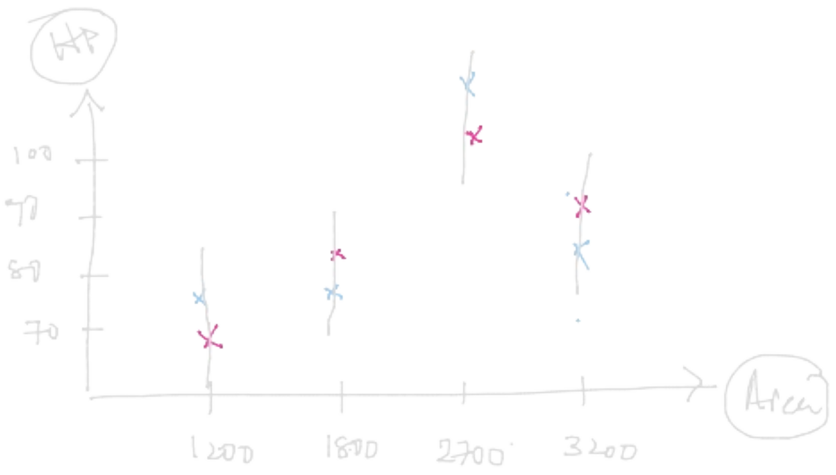
Errors

0	3.5
1	4.3
2	10.75 ←
3	1.3
4	6.4
5	13.75 ←
6	4 3

np.where(model.error > 10)

= [2, 5]

	Area	loc	Bedrooms	HP	$\hat{HP}$	Errors
→ 1.	1200	5	2	70	75	−5
2	1800	20	2.5	85	81	+4
3.	3200	7	3	93	89	+4
4	2700	10	3	102	108	−6

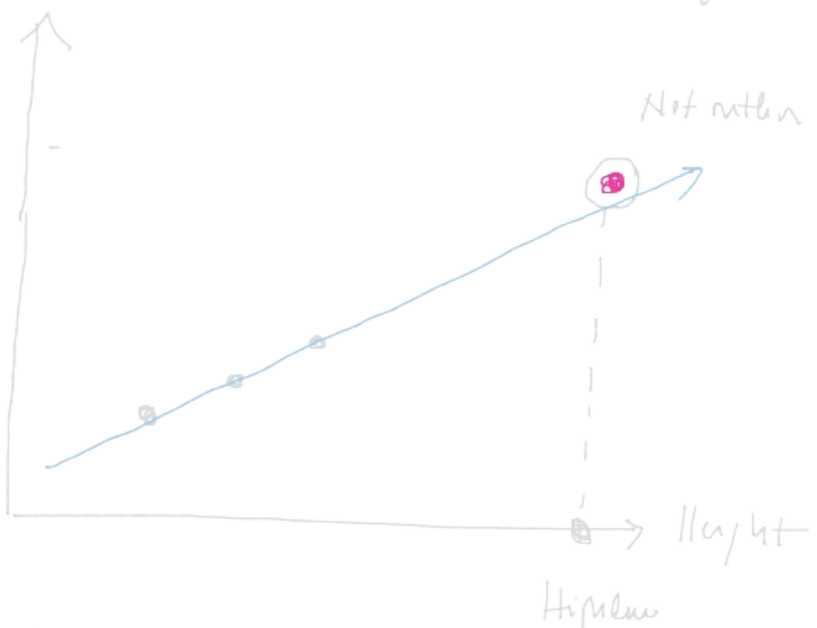


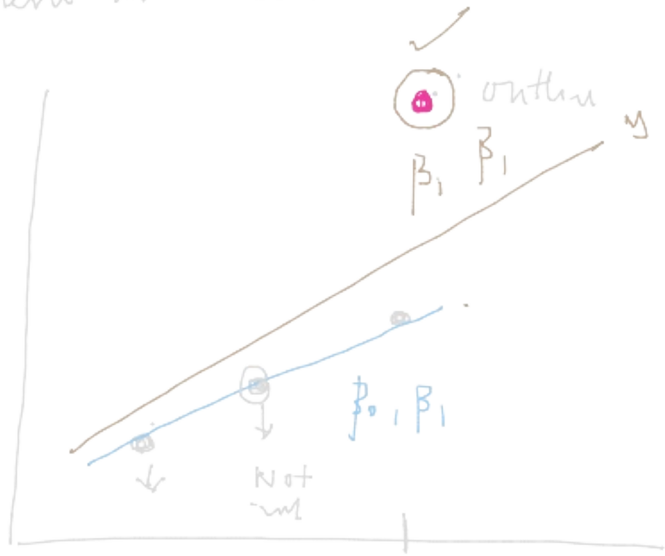
High.
Leverage Points

Extreme 'x'-Values.

Outliers

— Extreme Values for 'y'
— It does not follow the pattern in the data.





Highly Influential Points

— Changing the model itself

→ Highly Infl

Not influ

→ Highly Influ

Include P1 find β_0, β_1

Exclude P1 find β_0, β_1

Influence of a Point

'Cook's distance'

P1 → > 1.0 Highly influential

x_1	x_2	x_3	x_4	y
P1 → (				)
→ (				)

Build model with all records → $(\beta_0\ \beta_1)$ → reference

Remove (P+1) Build model → $(\beta_0\ \beta_1)$

Remove P2, Build the mdl → β_0, β_1

x_1 x_2 x_3 x_{5n} y

1 — — — — — High Influ C_1 ✓

C_2 ✗

$\beta_0, \beta_1 \rightarrow C_3$ ✗

$\beta_1, \beta_1 \rightarrow$

Cook's distance ⇒ > 1.0 ✗

1 → C_1 ✗ 1.0

2 →

3

8

1. Problems in Features (Colums) → W7 shmld be removed
2. Problems in Errors → 0, 76 are problematic.