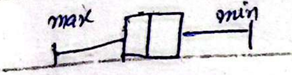
# Lecture 10 Data Preparation

جب ہم اپنے data کو اپنے مشین لرننگ ماڈل کے مطابق بناتے ہیں ،
اس عمل کو Data Preprocessing کہتے ہیں ۔ اس میں ہم اپنے
data set میں سے جو data ہمیں چاہیے ہوتا ہے صرف وہ لگاتے ہیں ۔

## Steps : Preprocessing :

M2 Huber SSE کا طریقہ :

Imp 1) Remove the outliers : اس میں ہم ایک error threshold define کرتے ہیں جس سے زیادہ error آئے اس کو ignore نہیں کرتے
Problem ? اگر remove نہ کریں تو ہمارا ماڈل outlier کی طرف move کرنا شروع کر دیتا ہے ۔



M1 (2 box Plot سے ہم اپنے data سے outlier remove کر سکتے ہیں ۔
جو بھی value max اور min سے باہر ہے اس کو remove کر دیں گے ۔

## 2) Missing Data :

1) جہاں جہاں missing ویلیو ہوگی وہ پوری کی پوری row remove کر دو ۔ (Inf. loss) ہوتا ہے
2) جہاں جس column میں missing ویلیو آ رہی ہے تو اس کالم کا average لے کر جہاں پر missing ویلیو ہے average کے ساتھ fill کر دیں ،
3) اپنے data کالم میں دیکھیں جو value سب سے زیادہ آ رہی ہو اس کو missing کی جگہ fill کر دیں ۔

## 3) Non-Numric Value handling to numric : (one hot Ecording)

1) اس میں ہم one hot Ecoding کرتے ہیں ۔ اس میں ہم کالم میں سے Unique ویلیوز کو final کرتے ہیں ۔ اگر unique کا سائز جتنا ہے اتنا matric بنانا ہے ۔ جیسے

$$\begin{array}{cccc} \text{Matric} & \text{Inter} & \text{Undergr} & \text{grad} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}$$

Matric ↓ replace
$[1\ 0\ 0\ 0]$
Inter ↓
$[0\ 1\ 0\ 0]$

جس جگہ وہ non-numeric value ہوگی اس کی جگہ Vector کو replace کر دیں گے ۔

جس جگہ پر one ہوگا جیسے first vector میں سب سے پہلے
one آ رہا ہے اس رو ہی میں وہ matrix سے belong کر رہا ہے۔

⇒ One More Method:

اس میں ہم اپنی تمام unique value کو کوئی نمبر سے assign کر دیں گے۔

Red → 1

Green → 2

Yellow → 3

again Red → 1

اس میں مسئلہ ہے کہ ہمارا ماڈل Sequence کو learn کرنا شروع کر دیتا ہے۔

4) Feature Scaling: (اس کو پیچھے لکھا ہوا ہے)

5) 5 number Summary: Box Plot

Steps:

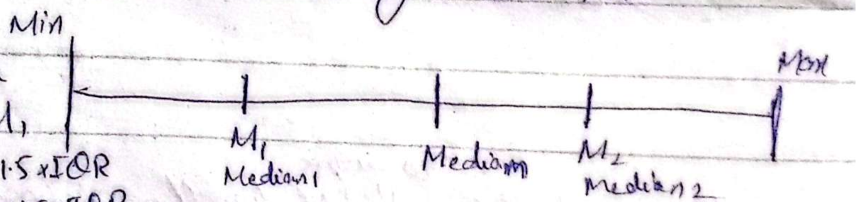
1) Sort dataset

2) Select Min & Max

3) IQR = $Q_2 - M_1$

4) Lower Limit = $Q_1 - 1.5 \times IQR$

5) Upper Limit = $Q_3 + 1.5 \times IQR$



⇒ Inter Qutial Range: IQR:

اس میں یہ پتا چلتا ہے کہ ہمارا زیادہ تر data کس جگہ پر ہے

6) Nois InJection:

جب ہمارا پاس data کا سائز کم ہو تو اس dataset کو Increse
کرنے کے method کو Nois InJection بولتے ہیں۔ اس میں
ہم original data میں تھوڑا تھوڑا تبدیل کر کے data میں شامل کر دیتے ہیں

⇒ Scaling:

اس میں ہم data کے column / attribute کو 1 سے 0 کی range میں لے کر آتے ہیں۔ اس میں ہم ہر attribute میں سب سے high value کو select کرتے ہیں اور ہر instance کو اس maximum سے divide کرتے ہیں اس کا فائدہ یہ ہے کہ training time کو کم کر دیتا ہے کرتے ہیں اس میں ہم کم iteration کرتے training کے لیے۔

Exp:

| Size (house | Scaling Size |
|---|---|
| 2000 | 2000 / 6000 = |
| 4000 | 4000 / 6000 = |
| max ↙ 6000 | 6000 / 6000 = |

Formula: $\text{Scaling} = \frac{\text{Size}(x)}{\max(\text{size})}$

Feature Scaling Method:

1) Mean Normalization:

1) اس میں سب سے پہلے column wise اس کا mean calculate کرتے ہیں۔

2) ہر value میں سے mean کو منفی کرنا ہے

3) پھر اس: column کے maximum سے divid کر دیں گے۔

$$\text{Normalization} = \frac{\overset{\text{size}}{\text{value}} - \overset{\text{size column}}{\text{mean}}}{\max}$$

2) Min - Max Scaling:

$$\underset{\text{Scaled feature}}{x_{new}} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3) Standard deviation Scaling:

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{n}}$$

# Box Plot & Five Number Summary

Dataset = [3,10,14,19,22,29,32,36,49,70]

Step1 : Sort the dataset

Dataset Sorted = [3, 10,14,19,22,29,32,36,49, $\boxed{70}$]

outlier

Step2 : Five number Summary

1- Minimum Value

2) First Quatial $Q_1$ (25%)

3) Median

4) Thrid Qustial $Q_3$ (75%)

5) Maximam.

Step3 :

1) Upper limit = $Q_3 + 1.5\,(IQR)$

$= 36 + 1.5(22)$ $\boxed{= \text{[illegible]} \;\; 69}$

2) Lower limit = $Q_1 - 1.5(IQR)$

$= 14 - 1.5(22)$ $\boxed{= \text{[illegible]} \;\; -19}$

Step 4 : Calculte $Q_1$ , $Q_3$ , IQR

1) $Q_1 = \frac{25}{100}$ (Total Instance + 1)

$= \frac{25}{100}(10+1) = \frac{275}{100}$ $\boxed{= 2.75}$

$\boxed{Q_1 = 2.75}$ → replace $\boxed{14}$

2) $Q_3 = \frac{75}{100}$ (total instance + 1)

$Q_3 = \frac{75}{100}(10+1) = \frac{825}{100} = 8.25$

$\boxed{Q_3 = 8.25}$ ⟹ replace [illegible] value $\boxed{Q_3 = 36}$

3) $IQR = Q_3 - Q_1$

IQR = 36 − 14

$\boxed{IQR = 22}$

⇒ Draw box Plot :

Dataset = $[3, 10, 14, 19, 22, 29, 32, 36, 49, \boxed{79}]$ outlier

Five number Sumary :

1) Minimum = 3

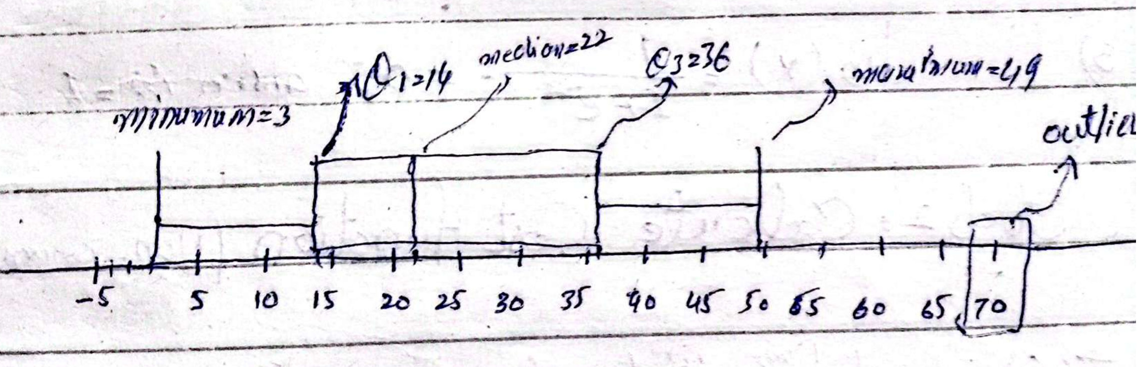
2) First Quartial $Q_1$ = 14

3) Median = 22

4) Third Quartial $Q_3$ = 36

5) Maximum = 49

Draw box Plot


minimum=3
$Q_1$=14
median=22
$Q_3$=36
maximum=49
outlier