NAME: ABU-BAKAR SIDDIQUE
SECTION: ~~23~~ BDS-3B
ROLL NO: 23L-2522
SUBMITTED TO: Dr. ASMA

1: DATA MUNGING:

(a) ANAMOLIES IN THE DATASET:

• **NAME:**
ID 5 contains name "123" which is not possible so, IT IS A ANOMALY.

• AGE:
ID 3 HAS AGE AS "FORTY FIVE" IT SHOULD BE 45 AS ALL ENTRIES ARE IN INTEGER FORM, ALSO AGE CANNOT BE NEGATIVE SO THAT -32 FOR ID 6 IS AN ANOMALY

• SALARY:
INSTANCES OF FEATURE SALARY IS INCONSISTENT i.e DIFFERANT FORMAT FOR EVERY INSTANCE.

• GENDER:
DIFFERENT FORMAT BEING OBSERVED THROUGHOUT THE FEATURE ALSO THERE IS A GENDER 123 WHICH IS NOT POSSIBLE

• JOIN DATE:

JOIN DATE IS CONTAINING INCONSIST DATA (W) IT SHOULD FOLLOW ONLY ONE FORMA EITHER "dd-mm-yyyy" or "mm-dd-yyyy".

• EXPERIENCE:

AT ID NUMBER ~~EIGTH~~ 8 IT IS SHOWIN EXPERIENCE "five" BUT OTHER EXPERIENCES ARE IN INTEGER, SO IT IS AN ANOMALY. ALSO THE UN OF EXPERIENCE LIKE YEAR OR MONTH IS'NT MENTIONED

• BONUS:

THE CURRENCY IN BONUS FEATURE IS NOT MENTIONED WHETHER POUND OR DOLLAR.

• NOTE:

EVERY "NAN" PRESENT IN THE feature-Vector NEEDS TO BE SORTED OUT, AND IS TREATED AS A ANOMALY

(b) D

• Dat

ieTHE

• HAN

PRE

(b) DATA MUNGING / WRANGLING STEPS :

• Data collection:
CHECKING IF DATA IS FROM RELIABLE SOURCES ie THE GIVEN DATASET.

• HANDLING MISSING DATA:
TREATING MISSING DATA IN A WAY THAT PRESERVES INTEGRITY OF THE ANALYSIS like Imputation.

- Mean
- Median
- Mode.

TRY TO REMOVE AS MUCH FLAGGING AS YOU CAN.

• HANDLING OUTLIERS:
TREATING OUTLIERS IN MY DATASET, FEATURES LIKE AGE, DATE (INSTANCES) WHICH CAN SKEW OUR ANALYSIS

• DATATYPE-CONVERSION:
ENSURING THAT DATATYPES IN MY DATASET ARE CONSISTENT AND APPROPRIATE.

• DATA BINNING / DISCRETIZATION:
SUCH STEPS TO BE TAKEN TO REDUCE NOISE FROM THE DATA.

c) HANDLING MISSING INSTANCES:

ID "~~8~~ 7" TUPLE SHOULD BE REMOVED BECAUSE IT HAS MANY MISSING INSTANCES OF EVERY FE

• AGE:

NAN FOR DIANA CAN COMPUTED BY TAKIN MEAN OF ALL FEMALES WHICH WOULD BE ≈31 SIMILARLY NAN FOR GEORGE CAN BE COMPUTED BY TAKING MEAN OF ALL MALES WHICH ≈ BUT WE DROPPING GEORGE TUPLE SO, IT DOE REQUIRE TAKING MEAN.

• SALARY:

JENNY SALARY CAN BE COMPUTED BY TAKING OF EVERY EMPLOYEE'S SALARIES WHICH WOULD BE ≈650 BECAUSE WE CAN SEE A RANGE OF MAJORITY B/W 60

• JOINING DATE:

AS THE GIVEN FEATURE IS NOT IN A O SO FAY'S JOINING DATE CAN BE COMPUTED BY CHECKI THE ORIGNAL DATA.

• DEPARTMENT:

ID "5" HAS SALARY 40K WHICH IS IN THE DOMAINS OF HR SO HE MIGHT BE A HR

ISSAC IS THE LOWEST PAID HE MIGHT BE A JANITOR OR ~~to~~ IN A LOWER CLASS DEPARTM

• EXPERIENCE:
BOB IS AN ENGINEAR WITH 70000$ PAY SO WE CAN COMPARE PEOPLE WHO ARE ENGINNER AND OUR EARNING B/W 65K TO 75K, HIS EXP MUST BE AROUND 5 YEARS.
DIANA IS A HR WITH 65K SALARY SEEING OTHER HR'S ATIHIS PAY HER EXP SHOULD BE ~~AROUND~~ MORE FIVE YEARS.

• BONUS:
WE SHOULD DROP BONUS FEATURE AS COMPUTING MEAN FOR SO MANY MISSING INSTANCES CAN SKEW OUR ANALYSIS.

QUESTION-02

DESCRIPTIVE STATISTICS

(a) MEAN MATCHES PLAYED BY PLAYERS

$$\text{Mean} = \frac{\sum \text{Matches PLAYES}}{\text{no. of PLAYERS}} \boxed{\approx 18.14 \approx 18} \text{ MATCHES}$$

(b) MEAN GOALS:

$\boxed{\text{Mean} = 7 \text{ GOALS}}$

~~ARE~~ MORE THAN 7 GOALS:
ALEX, LIAM, DVID
THREE PLAYERS WILL SCORE MORE MEAN GOALS

c) MEDIAN:
SORTING IN ORDER
4, 5, 6, $\boxed{7}$, 8, 9, 10 ⇒ 7 IS MEDIAN OF THE GOAL.

d) MODE:

THERE IS NO MODE AS NO NUMBER OF GOALS IS OCCURING MORE THAN ONCE.

e) CALCULATING S.D AND VARIANCE:

$$\text{VARIANCE} = \frac{\sum_{i=1}^{n} (x - x_i)^2}{n-1} \qquad \bar{x} = 7$$

$$\text{S.D} = \sqrt{\frac{\sum_{i=1}^{n} (x - x_i)^2}{n-1}} \qquad \therefore n - 1 = 6.$$

$(x - x_i)$	$(x - x_i)^2$
-2	4
1	1
0	0
3	9
-3	9
2	4
-1	1

$$\sum (x - x_i)^2 = 28.$$

$$\boxed{\text{VARIANCE} = 28/6 = 4.67.}$$

$$\boxed{\text{S.D}(x) = \sqrt{\text{VARIANCE}} = \sqrt{4.67} = 2.16}$$

(f) INFE
THE S.
IN TH
INCONSI
SCORING
PERFORM

(g) CO-
GOAL
$\sum (x-$
$\sum (x$
$\therefore$ CO-

$(y - y_i)$
-3.14
1.86
-0.14
3.8
-2.
0.
-1.

$\Rightarrow$ C

(f) INFER TO CONSISTENCY:

THE S.D OF 2.16 INDICATES A MODERATE CONSISTENCY IN THE PLAYER'S PERFORMANCES. THEY ARE NOT WIDELY INCONSISTENT, BUT THERE IS ENOUGH VARIATION IN GOAL SCORING FROM GAME TO GAME WHICH SUGGESTS THEIR PERFORMANCE IS NOT UNIFORM.

(g) CO-RELATION AND COVARIANCE:

GOALS

$\Sigma(x - x_i) =$

$\Sigma(x - x_i)^2 = 28$

MATCHES PLAYED

$\Sigma(y - y_i) = 0.02$

$\Sigma(y - y_i)^2 = 34.8572$

$\bar{x} = 7$

$\bar{y} = 18.14$

∴ CO-RELATION:

$$r = \frac{\Sigma(x_i - \bar{x}_n)(y_i - \bar{y}_q)}{\sqrt{\Sigma(x_i - \bar{x}_i)^2 \Sigma(y_i - \bar{y}_i)^2}}$$

$(y - y_i)$	$(y - y_i)^2$	$x - x_i$
-3.14	9.8596	-2
1.86	3.4596	1
-0.14	0.0196	0
3.86	14.8996	3
-2.14	4.5796	-3
0.86	0.7396	2
-1.14	1.2996	-1

~~$\Sigma x - x_i = (5-7) + (8-7) + (7-7) + (10-7) + (4-7) + (9-7) + (6-7)$~~

~~$= (-2) + (+1) + 0$~~

~~$= +3 + (-3) + 2 + (-1)$~~

~~$= 0$~~

$\Rightarrow (x - x_i)(y - y_i)$

6.28

1.86

0

11.58

6.42

1.72

1.14

NOTE: BOTH FEATURES HAVE ASSOCIATION AMONG EACH OTHER AS CO-RELATION ≈ 1

$\boxed{\Sigma(x - x_i)(y - y_i) = 29}$

$$co(x, y) = \frac{29}{\sqrt{28 \times (34.8572)}}$$

$\boxed{co(x, y) = 0.92}$

CO-VARIANCE:

$$\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{(29)}{6} = 4.83$$

NOTE: CO 4.83>0 SHOWS THE BOTH FEATURE HAVE LINEAR RELATIONSHIP AND THE MAGNITUDE SHOWS ITS STRENGTH UNLIKE CORELATION, CO-VARIANCE HAS NO RANGE

QUESTION-03:

(a) SKEWNESS FOR SUBJECT 1 MARKS:

$$\text{SKEWNESS} = \frac{1}{n}\left(\left[\sum \left(\frac{x_i-\bar{x}}{SD(x)}\right)\right]^3\right)$$

$$\bar{x} = \frac{\sum \text{MARKS}}{\text{no. of STUDENTS}} = \frac{555}{7} = 79.28$$

$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$\left(\frac{x-x_i}{SD}\right)$	$\left(\frac{x-x_i}{SD}\right)^3$	$\left(\frac{x-x_i}{SD}\right.$
5.72	32.71	0.58	0.19	0.1
-4.28	18.31	-0.4358	-0.082	0.0
12.72	161.79	1.29	2.14	2.07
-14.28	203.91	-1.45	-3.04	4.42
8.72	76	0.88	0.68	0.5
0.72	0.5184	0.073	3.8×10^{-4}	2.83
-9.28	86.11	-0.94	-0.83	0.7

$\sum (x_i-\bar{x}) = 0.04$

$\sum (x_i-\bar{x})^2 = 579.3484$

$\sum \left(\frac{x_i-\bar{x}}{SD(x)}\right)^3 = -0.94162$

$$SD(x) = \sqrt{\frac{579.34}{6}} \approx 9.82$$

$$\text{SKEWNESS} = \frac{1}{7}\left[\left(\frac{\cancel{0.04}}{\cancel{9.82}}\right)^3\right] \quad \frac{1}{7}(-0.941$$

SKEWNESS = −0.1345

• INTERPRETATION:

It IS SLIGHTLY

MEAN, MEDIAN,

MEAN < M

b) KURTOSIS:

$K = \frac{1}{n}\sum$

$\sum\left(\frac{X_i-\bar{X}}{SD}\right)^4 =$

$\boxed{K = \frac{1}{7}(8}$

EXCESS KUR

• INTEPRETATI

KURTOSIS

MEANING

PEAK COMP

EXCESS KUR

// //

// //

THIRD ON

ARE

CONCEN

A LESS

DEVIAT

• INTERPRETATION:

It IS SLIGHTLY NEGATIVELY SKEWED. RELATION B/W MEAN, MEDIAN, MODE WILL BE OBSERVED AS.

MEAN < MEADIAN < MODE.

b) KURTOSIS:

$$K = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{SD} \right)^4$$

$$\sum \left(\frac{x_i - \bar{x}}{SD} \right)^4 = 8.69.$$

$$\boxed{K = \tfrac{1}{7}(8.69) = 1.24}$$

EXCESS KURTOSIS => ~~3 7 K~~ = K-3 = -1.76

• INTEPRETATION

~~KURTOSIS~~ >0 MEANS ~~DISTRIBUTION IS~~ LEPTOKURTIC MEANING ~~DATA HAS~~ ~~HEAVY~~ TAILS AND A SHARPER PEAK COMPARED TO NORMAL DISTRIBUTION.

EXCESS KURTOSIS = 0 (NORMAL DISTRIBUTION)

" " >0 (HEAVY TAIL AND SHARPER PEAK).

" " <0 (LIGHT TAIL AND FLATTERED PEAK)

THIRD ONE IS OUR CASE WHICH MEANS: OUTLIERS ARE LESS FREQUENT AND DATA IS MORE CONCENTRATED AROUND THE MEAN BUT WITH A LESS PRONOUNCED PEAK AND FEWER EXTREME DEVIATIONS.

KURTOSIS FOR SUBJECT 2:

$$K = \frac{1}{n} \sum \left[\left(\frac{x_i - \bar{x}}{SD(x)} \right) \right]^4$$

$$\bar{x} = \frac{560}{7} = 80.$$

$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{x_i - \bar{x}}{SD}$	$\left(\frac{x_i - \bar{x}}{SD}\right)^4$
10	100	0.88	0.599
2	4	0.17	8.35×10^{-4}
15	225	1.32	3.0
-10	100	-0.88	0.599
7	49	0.61	0.138
-4	16	-0.35	0.015
-20	400	1.76	9.59

$\sum (x_i - \bar{x})^2 = 894$

$\boxed{SD(x) = \sqrt{\frac{894}{7}} = 11.30}$ $\therefore SD(x) = \sum \frac{(x - \bar{x}_1)^2}{n-1}$

$K_2 = \frac{1}{7}\left(13.94\right) = 1.99$

$\boxed{\text{EXCESS } K_2 = 1.99 - 3 = -1.00}$

SAME INTERPRETATION AS FOR K_1

KURTOSIS FOR SUBJECT 3:

$$K = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{SD(x)} \right)^4$$

$$\bar{x} = \frac{577}{7} = 82.42.$$

$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$\frac{x_i - \bar{x}}{SD}$	$\left[\frac{x_i - \bar{x}}{SD(x)}\right]^4$
-4.42	19.5	-0.52	0.073
5.58	31.13	0.65	0.17
8.58	73.61	1.00	1
-10.42	108.57	-1.22	2.21
11.58	134.09	1.36	3.42
-3.42	11.69	-0.4023	0.026
-7.42	55.05	-0.8729	0.5805

$$\sum (x_i - \bar{x})^2 = 433.64$$

$$SD(x) = \sqrt{\frac{433.64}{6}} \qquad \therefore SD(x) = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$SD(x) = 8.5$$

$$\sum \left(\frac{x_i - \bar{x}}{SD(x)} \right)^4 = 7.4795$$

$$K_3 = \frac{1}{7} (7.4795) = 1.06$$

$$\text{EXCESS } K_3 = 1.06 - 3 = -1.9315.$$

EXCESS $K < 0$

DATA ~~IS~~ HAS LIGHTER TAIL AND FLATTERED PEAK, A LESS PRONOUNCED PEAK AND FEWER EXTREME DEVIATIONS.

c) NORMALIZE SUBJECT-2 MARKS

USING Z-SCORE NORMALIZATION

$$\text{Z-SCORE NORMALIZATION} = \frac{X - \bar{X}}{SD(X)}$$

SUBJECT2	Z-SCORE
90	$\left(\frac{90-80}{11.3}\right) = 0.88$
82	$\left(\frac{82-80}{11.3}\right) = 0.17$
95	$(95-80)/11.3 = 1.32$
70	$(70-80)/11.3 = -0.88$
87	$(87-80)/11.3 = 0.61$
76	$(76-80)/11.3 = -0.35$
60	$(60-80)/11.3 = 1.76$

$\boxed{\bar{X} = 80}$

SD FOR SUBJECT2 = $\boxed{11.30}$

~~(X1×4)~~

$$Z = \frac{X_i - \bar{X}}{SD(X)}$$

IMPLEMENTING THIS FORMULA ON EACH MARKS OF A STUDENT AND WRITTING IN A NEW COLUMN.

NOTE:

Z-SCORE = 0 => DATA POINT IS EXACTLY AT ME

Z-SCORE > 0 => DATA POINT IS ABOVE TH

Z-SCORE < 0 => DATA POINT IS BELOW T MEAN

QUESTION

10, 1

a> EQU

BIN 1

BIN 2

BIN 3:

~~BIN 4:~~

b> SMOO

BIN 1

BIN 2

BIN

c> ME

BIN 1

BIN 2

BIN 3

QUESTION-04

10, 15, 22, 24, 30, 35, 40, 41, 42, 43, 45, 50.

a) EQUI-DEPTH BINNING:

BIN 1: 10, 15, 22, 24
BIN 2: 30, 35, 40, 41
BIN 3: 42, 43, 45, 50
~~BIN 4:~~

b) SMOOTHED VALUE FOR EACH BIN

BIN 1: 10, 10, 24, 24
BIN 2: 30, 30, 41, 41
BIN 3: 42, 42, 42, 50

c) MEAN OF SMOOTHED DATA BINS:

BIN 1: 18, 18, 18, 18 ∴ Mean ≈ 17.75
BIN 2: 37, 37, 37, 37 ∴ Mean ≈ 36.5
BIN 3: 45, 45, 45, 45. ∴ Mean ≈ 45

QUESTION-05

(a) ENTROPY AND INFO GAIN:

$\therefore$ ENTROPY $= -\sum P_{ij} \log_2 P_{ij}$

DATA SET:

A : 4, 6, 12, 16, 20

B : 8, 10, 14, 18, 22

Total = 10.

A = 5

B = 5

ENTROPY $= \frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}$

$= -\left(-\frac{1}{2} + \left(-\frac{1}{2}\right)\right) = -(-1) = 1$

• INFO GAIN W.R.T CLASS LABEL.

$E = \frac{5}{5}\log_2 \frac{5}{5} = 0$ } SAME FOR A AND B

No info Gain.

b) $E_{(D \leq 13)}$

A : 4, 6, 12

B : 8, 10

$E = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right)$

$-(-0.44 + (-0.528))$

$E = +0.96$

INFOGAIN 1 $= E_{TOTAL} - E_{SPLIT1}$

$\boxed{= 1 - 0.96 = 0.04}$

SPLIT 2 :

$E_{D \geq 13}$

A :

B

E =

WEIGH

Weighted

INFO

SPLIT 2:

$E_{D \geq 13}$.

A : 16 , 20

B 14 , 18 , 22

$E = 0.96$

WEIGHTED ENTROPY:

$\frac{5}{10}(0.96) + \frac{5}{10}(0.96)$

Weighted = 0.96

INFO GAIN = 1 − 0.96 = 0.04.

ME FOR

ND B

QUESTION:06

ENCODING GIVEN DATASET:

SIZE CATEGORY	PURCHASED
1	1
2	0
3	1
1	1
2	0
3	1
2	0
1	1
~~4~~3	0
2	1

KEY

1 = SMALL
2 = MEDIUM
3 = LARGE

0 = NO
1 = YES

QUESTION-07:

FORMULAS:

- $IQR = Q_3 - Q_1$
- $\text{LOWER BOUND} = Q_1 - 1.5 \times IQR$
- $\text{UPPER BOUND} = Q_3 + 1.5 \times IQR$.

i) FIRST STEP IS TO SORT THE DATASET.

(a) IDENTIFYING OUTLIERS USING IQR METHOD.

SORTED ON NEXT PAGE.

SALARY:

1) 40,0
2) 45,0
3) 55,0
4) 60,0
5) 85,0
6) 120.
7) 130.
8) 200,
9) 250
10) 300.

$Q_3 = 2$
$Q_1 =$
$Q_3 - Q$
$\boxed{IQR =}$

LOW LO
U

L.B
U.B

∴ ANY
CONS

EVER
WI
RANG
HAS

KEY

1 = SMALL
2 = MEDIUM
3 = LARGE

0 = NO
1 = YES

SALARY: (SORTED ORTER):

1) 40,000
2) 45,000
3) 55,000
4) 60,000
5) 85,000
6) 120,000
7) 130,000
8) 200,000
9) 250,000
10) 300,000

$\therefore Q1$ POSITION $= \frac{n+1}{4}$

$\therefore Q2$ POSITION $= \frac{n+1}{2}$

$\therefore Q3$ POSITION $= 3 \times \frac{n+1}{4}$

$Q_1 = \frac{10+1}{4} = 11/4 = 2.75 \approx 3.$

$Q_3 = 3(11/4) = 33/4 \approx 8.25 \approx 9.$

$\therefore$ LOWER BOUND $= Q_1 - 1.5IQR$

$\therefore$ UPPER BOUND $= Q_3 + 1.5IQR.$

$Q_3 = 250,000$

$Q_1 = 55,000$

$Q_3 - Q_1 = IQR$

$\boxed{IQR = 195,000}$

~~LOW~~ LOWER BOUND $= 55,000 - 292500 = -237500$

UPPER BOUND $= 250,000 + 1.5(295,000) = 542500$

L.B = −237500
U.B = 542500

$\therefore$ ANY VALUE OUTSIDE THESE BOUNDS WILLBE CONSIDERED AS AN OUTLIER.

DATASET.

IQR METHOD.

EVERY INSTANCE OF FEATURE SALARY IS WITHIN (~~THIS DATAS~~) THE INTER QUARTILE RANGES SO THE DATASET FEATURE SALARY HAS NO OUTLZER INIT.

(b) THERE ARE'NT ANY OUTLIERS IN THIS FEATURE.

(c) FEATURING HAVING NO OUTLIER SO ITS MEAN BEFORE AND AFTER WILL REMAIN THE SAME BECAUSE IT CONTAINS NO OUTLIERS

$$MEAN = \frac{\sum SALARIES}{no.of\ employees} = \frac{12,85000}{10} \boxed{=128500}$$

(d) SINCE THERE WERE'NT ANY OUTLIERS SO THE DATASET WON'T BE AFFECTED. IN ANY TERMS.